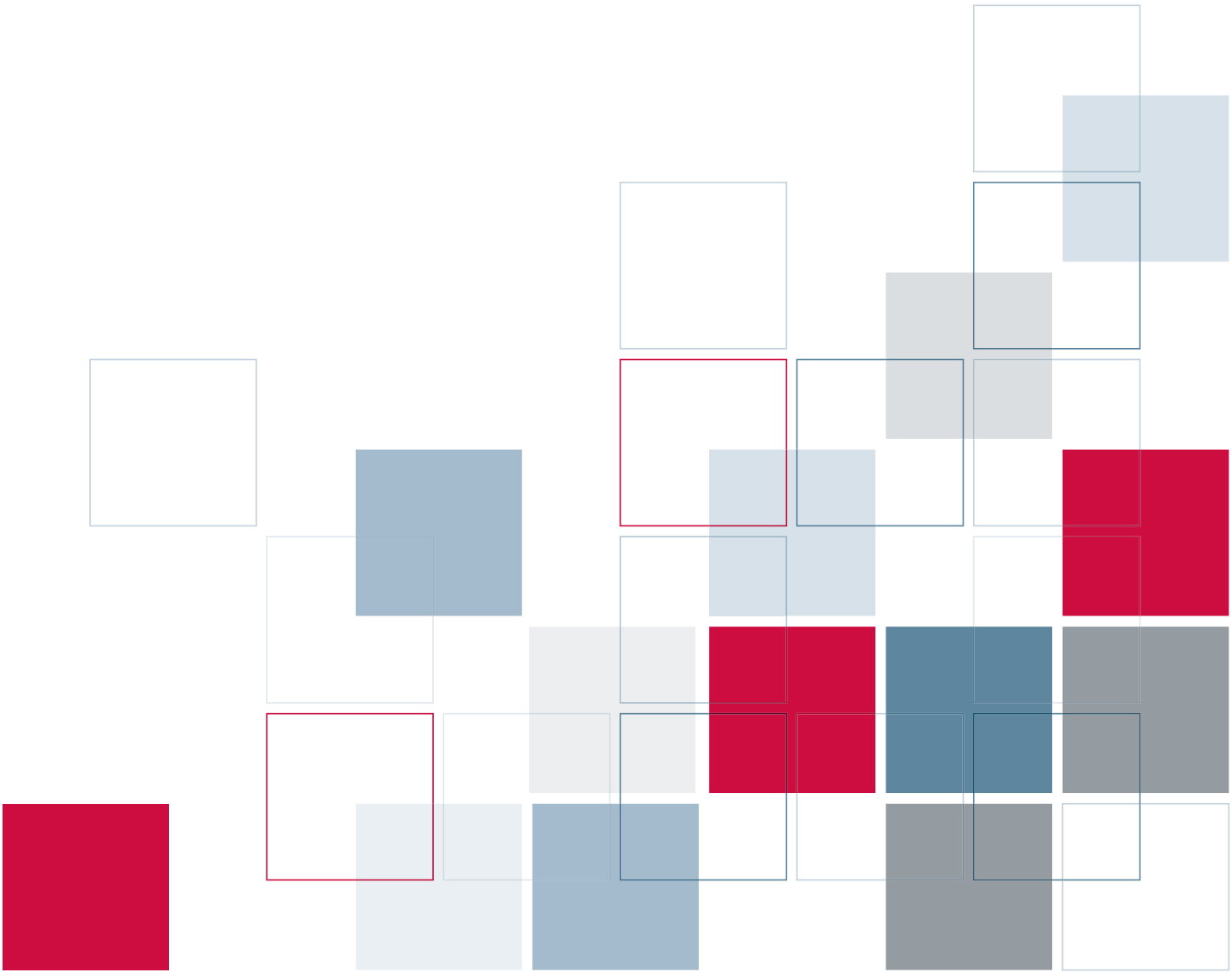
SPSS Categories™ 16.0



Weitere Informationen zu SPSS®-Software-Produkten finden Sie auf unserer Website unter der Adresse *http://www.spss.com* oder wenden Sie sich an

SPSS Inc.
233 South Wacker Drive, 11th Floor
Chicago, IL 60606-6412, USA
Tel.: (312) 651-3000
Fax: (312) 651-3668

SPSS ist eine eingetragene Marke, und weitere Produktnamen sind Marken der SPSS Inc. für Computerprogramme von SPSS Inc. Die Herstellung oder Verbreitung von Materialien, die diese Programme beschreiben, ist ohne die schriftliche Erlaubnis des Eigentümers der Marke und der Lizenzrechte der Software und der Copyrights der veröffentlichten Materialien verboten.

Die SOFTWARE und die Dokumentation werden mit BESCHRÄNKTEN RECHTEN zur Verfügung gestellt. Verwendung, Vervielfältigung und Veröffentlichung durch die Regierung unterliegen den Beschränkungen in Unterabschnitt (c)(1)(ii) von The Rights in Technical Data and Computer Software unter 52.227-7013. Vertragspartner/Hersteller ist SPSS Inc., 233 South Wacker Drive, 11th Floor, Chicago, IL 60606-6412.
Patentnr. 7.023.453

Allgemeiner Hinweis: Andere in diesem Dokument verwendete Produktnamen werden nur zu Identifikationszwecken genannt und können Marken der entsprechenden Unternehmen sein.

Windows ist eine eingetragene Marke der Microsoft Corporation.

Apple, Mac und das Mac-Logo sind Marken von Apple Computer, Inc., die in den USA und in anderen Ländern eingetragen sind.

Dieses Produkt verwendet WinWrap Basic, Copyright 1993–2007, Polar Engineering and Consulting, *http://www.winwrap.com*.

SPSS Categories™ 16.0
Copyright © 2007 SPSS Inc.
Alle Rechte vorbehalten.

Ohne schriftliche Erlaubnis der SPSS GmbH Software darf kein Teil dieses Handbuchs für irgendwelche Zwecke oder in irgendeiner Form mit irgendwelchen Mitteln, elektronisch oder mechanisch, mittels Fotokopie, durch Aufzeichnung oder durch andere Informationsspeicherungssysteme reproduziert werden.

Vorwort

SPSS 16.0 ist ein umfassendes System zum Analysieren von Daten. Das optionale Zusatzmodul Categories (Kategorien) bietet die zusätzlichen Analyseverfahren, die in diesem Handbuch beschrieben sind. Die Prozeduren im Zusatzmodul Categories (Kategorien) müssen zusammen mit SPSS 16.0 Base verwendet werden. Sie sind vollständig in dieses System integriert.

Installation

Zur Installation von SPSS Categories (Kategorien) Erweiterungsmodul führen Sie den Lizenzautorisierungsassistenten mit dem Autorisierungscode aus, den Sie von SPSS erhalten haben. Weitere Informationen finden Sie in den Installationsanweisungen im Lieferumfang von SPSS Categories (Kategorien) Erweiterungsmodul.

Kompatibilität

SPSS kann auf vielen Computersystemen ausgeführt werden. Mindestanforderungen an das System und Empfehlungen finden Sie in den Unterlagen, die mit Ihrem System geliefert werden.

Seriennummern

Die Seriennummer des Programms dient gleichzeitig als Identifikationsnummer bei SPSS. Sie benötigen diese Seriennummer, wenn Sie sich an SPSS wenden, um Informationen über Kundendienst, zu Zahlungen oder Aktualisierungen des Systems zu erhalten. Die Seriennummer wird mit dem Base-System ausgeliefert.

Kundendienst

Wenden Sie sich mit Fragen bezüglich der Lieferung oder Ihres Kundenkontos an Ihr regionales SPSS-Büro, das Sie auf der SPSS-Website unter *http://www.spss.com/worldwide* finden. Halten Sie bitte stets Ihre Seriennummer bereit.

Ausbildungsseminare

SPSS bietet öffentliche und unternehmensinterne Seminare an. Alle Seminare beinhalten auch praktische Übungen. Seminare finden in größeren Städten regelmäßig statt. Wenn Sie weitere Informationen zu diesen Schulungen wünschen, wenden Sie sich an Ihr regionales SPSS-Büro, das Sie auf der SPSS-Website unter *http://www.spss.com/worldwide* finden.

Technischer Support

Kunden von SPSS mit Wartungsvertrag können den Technischen Support in Anspruch nehmen. Kunden können sich an den Technischen Support wenden, wenn sie Hilfe bei der Arbeit mit SPSS oder bei der Installation in einer der unterstützten Hardware-Umgebungen benötigen. Informationen über den Technischen Support finden Sie auf der Website von SPSS unter *http://www.spss.com* oder wenden Sie sich an Ihr regionales SPSS-Büro, das Sie auf der SPSS-Website unter *http://www.spss.com/worldwide* finden. Bei einem Anruf werden Sie nach Ihrem Namen, dem Namen Ihrer Organisation und Ihrer Seriennummer gefragt.

Weitere Veröffentlichungen

Weitere Exemplare von Produkthandbüchern können direkt bei SPSS Inc. bestellt werden. Besuchen Sie den SPSS Web Store unter *http://www.spss.com/estore* oder wenden Sie sich an Ihr regionales SPSS-Büro, das Sie auf der SPSS-Website unter *http://www.spss.com/worldwide* finden. Wenden Sie sich bei telefonischen Bestellungen in den USA und Kanada unter 800-543-2185 direkt an SPSS Inc. Wenden Sie sich bei telefonischen Bestellungen außerhalb von Nordamerika an Ihr regionales SPSS-Büro, das Sie auf der SPPS-Website finden.

Das Handbuch *SPSS Statistical Procedures Companion* von Marija Norušis wurde von Prentice Hall veröffentlicht. Eine neue Fassung dieses Buchs mit Aktualisierungen für SPSS 16.0 ist geplant. Das Handbuch *SPSS Advanced Statistical Procedures Companion*, bei dem auch SPSS 16.0 berücksichtigt wird, erscheint demnächst. Das Handbuch *SPSS Guide to Data Analysis* für SPSS 16.0 wird ebenfalls derzeit erstellt. Ankündigungen für Veröffentlichungen, die ausschließlich über Prentice Hall verfügbar sind, finden Sie auf der SPSS-Website unter *http://www.spss.com/estore* (wählen Sie Ihr Land aus und klicken Sie auf Books).

Kundenmeinungen

Ihre Meinung ist uns wichtig. Teilen Sie uns bitte Ihre Erfahrungen mit SPSS-Produkten mit. Insbesondere haben wir Interesse an neuen, interessanten Anwendungsgebieten von SPSS Categories (Kategorien) Erweiterungsmodul. Senden Sie uns eine E-Mail an *suggest@spss.com* oder schreiben Sie an: SPSS Inc., Attn: Director of Product Planning, 233 South Wacker Drive, 11th Floor, Chicago, IL 60606-6412.

Über dieses Handbuch

In diesem Handbuch wird die grafische Benutzeroberfläche für die in SPSS Categories (Kategorien) Erweiterungsmodul enthaltenen Prozeduren erläutert. Die Abbildungen der Dialogfelder stammen aus SPSS. Detaillierte Informationen zur Befehlssyntax für die Funktionen in SPSS Categories (Kategorien) Erweiterungsmodul sind auf zwei Arten verfügbar: als Bestandteil des umfassenden Hilfesystems und als separates Dokument im PDF-Format im Handbuch *SPSS 16.0 Command Syntax Reference*, das auch über das Menü “Hilfe” verfügbar ist.

Kontakt zu SPSS

Wenn Sie in unseren Verteiler aufgenommen werden möchten, wenden Sie sich an eines unserer Büros, die Sie auf unserer Website unter *http://www.spss.com/worldwide* finden.

Danksagung

Die Prozeduren für die optimale Skalierung und ihre Implementierung in SPSS wurden von der Data Theory Scaling System Group (DTSS) entwickelt, die aus Mitgliedern der Institute für Pädagogik und Psychologie der sozialwissenschaftlichen Fakultät der Universität Leiden besteht.

Willem Heiser, Jacqueline Meulman, Gerda van den Berg und Patrick Groenen waren an den ursprünglichen Prozeduren aus dem Jahr 1990 beteiligt. Jacqueline Meulman und Peter Neufeglise waren an der Entwicklung von Prozeduren für die kategoriale Regression, die Korrespondenzanalyse, die kategoriale Hauptkomponentenanalyse und die multidimensionale Skalierung beteiligt. Außerdem leistete Anita van der Kooij Beiträge insbesondere zu CATREG, CORRESPONDENCE und CATPCA. Willem Heiser, Jacques Commandeur, Frank Busing, Gerda van den Berg und Patrick Groenen waren an der Entwicklung der Prozedur PROXSCAL beteiligt. Frank Busing, Willem Heiser, Patrick Groenen und Peter Neufeglise waren an der Entwicklung der Prozedur PROXSCAL beteiligt.

Inhalt

Teil I: Benutzerhandbuch

1 Einführung in die Prozeduren für die optimale Skalierung von kategorialen Daten 1

Was ist optimale Skalierung? . . . 1
Warum sollte optimale Skalierung verwendet werden? . . . 1
Messniveau der optimalen Skalierung . . . 2
Auswahl des Messniveaus der optimalen Skalierung . . . 3
Transformationsdiagramme . . . 3
Kategoriecodes . . . 4
Welche Prozedur ist für Ihre Anwendung am besten geeignet? . . . 7
Kategoriale Regression . . . 8
Kategoriale Hauptkomponentenanalyse . . . 9
Nichtlineare kanonische Korrelationsanalyse . . . 9
Korrespondenzanalyse . . . 10
Mehrfachkorrespondenzanalyse . . . 12
Multidimensionale Skalierung . . . 13
Multidimensionale Entfaltung . . . 13
Seitenverhältnis in Diagrammen mit optimaler Skalierung . . . 14
Empfohlene Literatur . . . 14

2 Kategoriale Regression (CATREG) 16

Definieren der Skala in der kategorialen Regression . . . 17
Kategoriale Regression: Diskretisierung . . . 19
Fehlende Werte in kategorialer Regression . . . 20
Kategoriale Regression: Optionen . . . 21
Kategoriale Regression: Ausgabe . . . 21
Kategoriale Regression: Speichern . . . 23
Kategoriale Regression: Transformationsdiagramme . . . 23
Zusätzliche Funktionen beim Befehl CATREG . . . 24

3 Kategoriale Hauptkomponentenanalyse (CATPCA) 25

Skala und Gewicht definieren in CATPCA . . . 27
Kategoriale Hauptkomponenten: Diskretisierung . . . 29
Kategoriale Hauptkomponenten: Fehlende Werte . . . 30
Kategoriale Hauptkomponenten: Optionen . . . 31
Kategoriale Hauptkomponenten: Ausgabe . . . 33
Kategoriale Hauptkomponenten: Speichern . . . 35
Kategoriale Hauptkomponenten: Objekt- und Variablendiagramme . . . 36
Kategoriale Hauptkomponenten: Kategoriendiagramme . . . 37
Kategoriale Hauptkomponentenanalyse: Ladungsdiagramme . . . 38
Zusätzliche Funktionen beim Befehl CATPCA . . . 39

4 Nichtlineare kanonische Korrelationsanalyse (OVERALS) 40

Bereich und Skala definieren . . . 43
Bereich definieren . . . 44
Nichtlineare kanonische Korrelationsanalyse: Optionen . . . 44
Zusätzliche Funktionen beim Befehl OVERALS . . . 45

5 Korrespondenzanalyse 47

Definieren des Zeilenbereichs bei der Korrespondenzanalyse . . . 49
Definieren des Spaltenbereichs in der Korrespondenzanalyse . . . 49
Korrespondenzanalyse: Modell . . . 50
Korrespondenzanalyse: Statistiken . . . 52
Korrespondenzanalyse: Diagramme . . . 53
Zusätzliche Funktionen beim Befehl CORRESPONDENCE . . . 55

6 Mehrfachkorrespondenzanalyse 56

Definition des Variablengewichts bei der Mehrfachkorrespondenzanalyse . . . 58
Mehrfachkorrespondenzanalyse: Diskretisierung . . . 59
Mehrfachkorrespondenzanalyse: Fehlende Werte . . . 60
Mehrfachkorrespondenzanalyse: Optionen . . . 61

Mehrfachkorrespondenzanalyse: Ausgabe 63
Mehrfachkorrespondenzanalyse: Speichern 64
Mehrfachkorrespondenzanalyse: Objektdiagramme 65
Mehrfachkorrespondenzanalyse: Variablendiagramme 66
Zusätzliche Funktionen beim Befehl MULTIPLE CORRESPONDENCE 67

7 Multidimensionale Skalierung (PROXSCAL) 69

Ähnlichkeiten in mehreren Spalten von Matrizen 72
Ähnlichkeiten in Spalten 73
Ähnlichkeiten in einer Spalte 74
Ähnlichkeiten aus Daten erstellen 75
Distanzen aus Daten erstellen 76
Definieren eines Modells für die multidimensionale Skalierung 77
Multidimensionale Skalierung: Einschränkungen 78
Multidimensionale Skalierung: Optionen 79
Multidimensionale Skalierung: Diagramme, Version 1 81
Multidimensionale Skalierung: Diagramme, Version 2 82
Multidimensionale Skalierung: Ausgabe 83
Zusätzliche Funktionen beim Befehl PROXSCAL 84

8 Multidimensionale Entfaltung (PREFSCAL) 85

Definieren eines Modells für die multidimensionale Entfaltung 87
Multidimensionale Entfaltung: Einschränkungen 88
Multidimensionale Entfaltung: Optionen 89
Multidimensionale Entfaltung: Diagramme 91
Multidimensionale Entfaltung: Ausgabe 93
Zusätzliche Funktionen beim Befehl PREFSCAL 95

Teil II: Beispiele

9 Kategoriale Regression 97

Beispiel: Daten für Teppichreiniger . . . 97
Standardverfahren der linearen Regressionsanalyse . . . 98
Kategoriale Regressionsanalyse . . . 104
Beispiel: Ozondaten . . . 117
Diskretisierung von Variablen . . . 118
Auswahl des Transformationstyps . . . 118
Optimalität der Quantifikationen . . . 128
Auswirkungen von Transformationen . . . 130
Empfohlene Literatur . . . 139

10 Kategoriale Hauptkomponentenanalyse 141

Beispiel: Untersuchung der Interrelationen sozialer Systeme . . . 141
Durchführung der Analyse . . . 142
Anzahl der Dimensionen . . . 147
Quantifizierungen . . . 148
Objektwerte . . . 150
Komponentenladungen . . . 151
Zusätzliche Dimensionen . . . 152
Beispiel: Symptomatologie von Eßstörungen . . . 154
Durchführung der Analyse . . . 156
Transformationsdiagramme . . . 168
Modellzusammenfassung . . . 170
Komponentenladungen . . . 171
Objektwerte . . . 172
Untersuchung der Strukturen im Krankheitsverlauf . . . 174
Empfohlene Literatur . . . 190

11 Nichtlineare kanonische Korrelationsanalyse 192

Beispiel: Analyse der Umfrageergebnisse . . . 192
Untersuchen der Daten . . . 193
Berücksichtigen der Ähnlichkeiten zwischen Sets . . . 199
Komponentenladungen . . . 203

Transformationsdiagramme 204
Einzelkategoriekoordinaten im Vergleich mit Mehrfachkategoriekoordinaten 206
Zentroide und projizierte Zentroide 207
Eine alternative Analyse 209
Allgemeine Vorschläge 214
Empfohlene Literatur 215

12 Korrespondenzanalyse 216

Normalisierung 217
Beispiel: Rauchgewohnheit nach Berufskategorie 218
Durchführung der Analyse 218
Korrespondenztabelle 221
Dimensionen 221
Biplot 222
Profile und Distanzen 223
Zeilen- und Spaltenwerte 225
Permutationen der Korrespondenztabelle 227
Konfidenzstatistiken 227
Zusatzprofile 228
Beispiel: Wahrnehmung von Kaffeemarken 234
Durchführung der Analyse 234
Dimensionen 238
Beiträge 239
Diagramme 241
Symmetrische Normalisierung 242
Beispiel: Flugmeilen zwischen Städten 244
Korrespondenztabelle 251
Zeilen- und Spaltenwerte 252
Empfohlene Literatur 253

13 Mehrfachkorrespondenzanalyse 254

Beispiel: Eigenschaften von Eisenwaren 254
Durchführung der Analyse 255
Modellzusammenfassung 258
Objektwerte 259
Diskriminationsmaße 260
Kategorienquantifikationen 261

Genauere Untersuchung der Objektwerte 263
Ausschluß von Ausreißern 266
Empfohlene Literatur 269

14 Multidimensionale Skalierung 271

Beispiel: Untersuchung von Bezeichnungen des Verwandtschaftsgrades 271
Festlegen der Anzahl der Dimensionen 272
Eine dreidimensionale Lösung 279
Dreidimensionale Lösung mit nicht standardmäßigen Transformationen 285
Diskussion 288
Empfohlene Literatur 288

15 Multidimensionale Entfaltung 290

Beispiel: Bevorzugte Frühstücksartikel 290
Erstellen einer degenerierten Lösung 290
Maße 293
Gemeinsamer Raum 294
Durchführen einer nicht degenerierten Analyse 295
Maße 296
Gemeinsamer Raum 297
Beispiel: Dreifach-Entfaltung von bevorzugten Frühstücksartikeln 297
Durchführung der Analyse 298
Maße 302
Gemeinsamer Raum 303
Individueller Raum 304
Verwenden einer anderen Ausgangskonfiguration 307
Maße 309
Gemeinsamer Raum 310
Individueller Raum 311
Beispiel: Untersuchen der Angemessenheit von Verhaltensweisen in bestimmten Situationen . . 313
Durchführung der Analyse 314
Maße 319
Gemeinsamer Raum 320
Ähnlichkeitstransformationen 321
Ändern der Ähnlichkeitstransformation (Ordinal) 321
Maße 323

Gemeinsamer Raum. 324
Ähnlichkeitstransformationen 325
Empfohlene Literatur 325

Anhang

A Beispieldateien *326*

Bibliografie *338*

Index *344*

Teil I: Benutzerhandbuch

Kapitel

1

Einführung in die Prozeduren für die optimale Skalierung von kategorialen Daten

Bei den Prozeduren in Categories wird die optimale Skalierung für die Analyse von Daten verwendet, die mit statistischen Standardprozeduren nur schwer oder gar nicht analysiert werden können. In diesem Kapitel werden die einzelnen Prozeduren, die Situationen, in denen die einzelnen Prozeduren am besten geeignet sind, die Beziehungen zwischen den Prozeduren und die Beziehungen der Prozeduren zu ihrem jeweiligen Gegenstück aus der Standardstatistik beschrieben.

Anmerkung: Diese Prozeduren und ihre Implementierung in SPSS wurden von der Data Theory Scaling System Group (DTSS) entwickelt, die aus Mitgliedern der Institute für Pädagogik und Psychologie der sozialwissenschaftlichen Fakultät der Universität Leiden besteht.

Was ist optimale Skalierung?

Der Sinn der optimalen Skalierung ist es, den Kategorien jeder Variablen numerische Quantifizierungen zuzuweisen und dadurch zu ermöglichen, dass Standardprozeduren verwendet werden können, um eine Lösung für die quantifizierten Variablen zu erhalten.

Die optimalen Skalenwerte werden den Kategorien der einzelnen Variablen anhand des Optimierungskriteriums für die verwendete Prozedur zugewiesen. Anders als die ursprünglichen Labels der nominalen oder ordinalen Variablen in der Analyse haben diese Skalenwerte metrische Eigenschaften.

In den meisten Prozeduren in Categories wird die optimale Quantifizierung für die einzelnen skalierten Variablen über eine iterative Methode, **alternierende kleinste Quadrate** genannt, erzielt, bei der zunächst die aktuellen Quantifizierungen verwendet werden, um eine Lösung zu finden, und dann die Quantifizierungen anhand dieser Lösung aktualisiert werden. Anschließend werden die aktualisierten Quantifizierungen verwendet, um eine neue Lösung zu finden, anhand derer wiederum die Quantifizierungen aktualisiert werden, usw. Dieser Vorgang wird fortgesetzt, bis ein Stopp-Kriterium erreicht wird.

Warum sollte optimale Skalierung verwendet werden?

Kategoriale Daten finden sich oft bei der Marktforschung, der Auswertung von Umfragen und in der Sozial- und Verhaltensforschung. Zahlreiche Forscher haben fast ausschließlich mit kategorialen Daten zu tun.

Es gibt zwar von den meisten Standardmodellen Anpassungen für die Analyse kategorialer Daten, doch diese funktionieren häufig nicht gut für Daten-Sets mit folgenden Eigenschaften:

- Zu wenige Beobachtungen
- Zu viele Variablen
- Zu viele Werte pro Variable

Bei der Quantifizierung von Kategorien können Verfahren zur optimalen Skalierung die Probleme in diesen Fällen vermeiden. Außerdem sind sie selbst dann nützlich, wenn spezialisierte Verfahren angebracht sind.

Anstatt auf der Interpretation von Parameterschätzern beruht die Interpretation der Ausgabe bei der optimalen Skalierung häufig auf grafischen Darstellungen. Verfahren zur optimalen Skalierung bieten hervorragende explorative Analysen, die eine gute Ergänzung zu anderen SPSS-Modellen darstellen. Durch Einengung des Fokus der Untersuchung kann die Visualisierung der Daten durch optimale Skalierung die Grundlage einer Analyse bilden, die sich auf die Interpretation von Modellparametern konzentriert.

Messniveau der optimalen Skalierung

Dieses Konzept kann bei der Verwendung der Prozeduren von Categories zunächst sehr verwirrend sein. Bei der Angabe des Niveaus geben Sie nicht das Niveau an, auf dem die Variablen *gemessen* werden, sondern das Niveau, auf dem sie *skaliert* werden. Der Grund dafür ist, dass die zu quantifizierenden Variablen möglicherweise nichtlineare Relationen aufweisen, unabhängig davon, wie sie gemessen werden.

Für Categories gibt es drei grundlegende Messniveaus:

- Beim **nominalen** Niveau wird davon ausgegangen, dass die Werte einer Variablen für nicht geordnete Kategorien stehen. Beispiele für möglicherweise nominale Variablen sind Bundesland, Postleitzahl, Religionszugehörigkeit und Multiple-Choice-Kategorien.
- Beim **ordinalen** Niveau wird davon ausgegangen, dass die Werte einer Variablen für geordnete Kategorien stehen. Beispiele sind Einstellungsskalen, die für den Grad der Zufriedenheit oder der Zuversicht stehen, sowie Werte für Präferenzen.
- Beim **numerischen** Niveau wird davon ausgegangen, dass die Werte einer Variablen für Kategorien mit einer sinnvollen Metrik stehen, sodass man sinnvolle Aussagen über die Abstände zwischen den Kategorien machen kann. Metrische Variablen sind beispielsweise Alter (in Jahren) oder Einkommen (in Geldeinheiten).

Nehmen wir beispielsweise an, die Variablen *Region*, *Tätigkeit* und *Alter* seien wie in der folgenden Tabelle kodiert.

Tabelle 1-1
Kodierungsschema für Region, Tätigkeit und Alter.

Region		**Tätigkeit**		**Alter**	
1	Norden	1	Praktikant	20	zwanzig Jahre alt
2	Süden	2	Vertreter	22	zweiundzwanzig Jahre alt
3	Osten	3	Manager	25	fünfundzwanzig Jahre alt
4	Westen			27	siebenundzwanzig Jahre alt

Die angezeigten Werte stehen für die Kategorien der einzelnen Variablen. *Region* ist eine nominale Variable. Die Variable *Region* weist vier Kategorien ohne implizierte Reihenfolge auf. Die Werte 1 bis 4 stehen einfach für die vier Kategorien; das Kodierungsschema ist völlig willkürlich. *Tätigkeit* dagegen könnte als ordinale Variable betrachtet werden. Die ursprünglichen Kategorien stellen eine Steigerung vom Praktikanten bis hin zum Manager dar. Höhere Codewerte stehen für eine Tätigkeit, die auf der innerbetrieblichen Karriereleiter weiter oben liegt. Es ist jedoch nur die Reihenfolge bekannt. Über den Abstand zwischen benachbarten Kategorien wird nichts ausgesagt. Im Gegensatz dazu kann *Alter* als numerische Variable betrachtet werden. Bei *Alter* haben die Abstände zwischen den Werten eine implizite Bedeutung. Der Abstand zwischen 20 und 22 ist genauso groß wie der Abstand zwischen 25 und 27, wohingegen der Abstand zwischen 22 und 25 größer als die beiden anderen ist.

Auswahl des Messniveaus der optimalen Skalierung

Es ist wichtig zu begreifen, dass es keine intrinsischen Eigenschaften einer Variablen gibt, die automatisch vordefinieren, welches Messniveau der optimalen Skalierung dafür angegeben werden sollte. Sie können die Daten so analysieren, wie es sinnvoll erscheint und die Interpretation erleichtert. Bei der Analyse einer Variablen mit eigentlich numerischem Niveau auf dem ordinalen Niveau kann durch die Verwendung einer nichtlinearen Transformation eine Lösung in weniger Dimensionen möglich sein.

Die folgenden zwei Beispiele zeigen, dass das naheliegendste Messniveau nicht unbedingt das beste Niveau für die optimale Skalierung sein muss. Angenommen, eine Variable ordnet Objekte in Altersgruppen ein. Alter kann zwar als numerische Variable skaliert werden, aber dennoch steht Sicherheit für Personen unter 25 Jahren möglicherweise in einer positiven Relation zum Alter, wohingegen sie für Personen über 60 Jahren in einer negativen Relation zum Alter steht. In diesem Fall kann es besser sein, das Alter als nominale Variable zu behandeln.

Zweites Beispiel: Eine Variable, die Personen nach politischen Neigungen ordnet, scheint ihrem Wesen nach nominal zu sein. Wenn Sie jedoch die Parteien von politisch links nach politisch rechts sortieren, wäre es sinnvoll, wenn die Quantifizierung der Parteien diese Reihenfolge berücksichtigen würde. Dazu müßte ein ordinales Niveau für die Analyse verwendet werden.

Auch wenn es keine vordefinierten Eigenschaften gibt, die eine Variable ausschließlich einem bestimmten Messniveau zuordnen, gibt es einige allgemeine Richtlinien, nach denen sich unerfahrene Benutzer richten können. Bei einzeln nominaler Quantifizierung kennen Sie normalerweise nicht die Reihenfolge der Kategorien, möchten jedoch, dass durch die Analyse eine festgelegt wird. Wenn die Reihenfolge der Kategorien bekannt ist, sollten Sie eine ordinale Quantifizierung verwenden. Wenn sich die Kategorien nicht ordnen lassen, sollten Sie eine mehrfach nominale Quantifizierung verwenden.

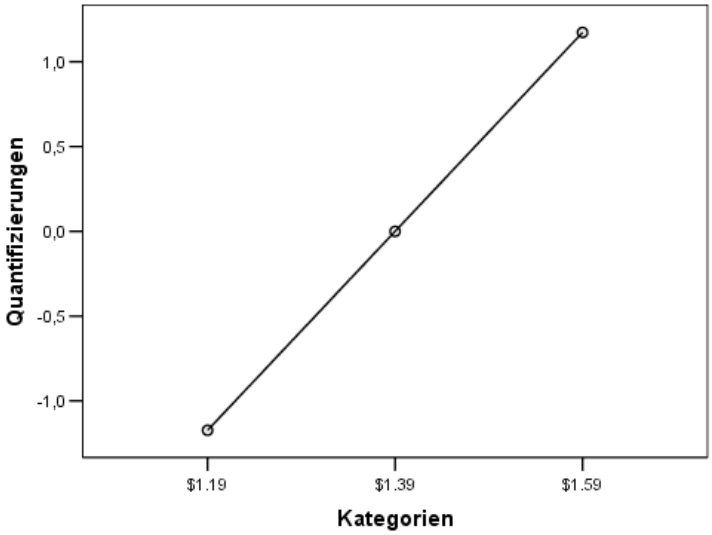
Transformationsdiagramme

Die verschiedenen Niveaus, auf denen die einzelnen Variablen skaliert werden können, bringen verschiedene Einschränkungen für die Quantifizierungen mit sich. Transformationsdiagramme illustrieren die Beziehung zwischen den Quantifizierungen und den ursprünglichen Kategorien, die sich aus dem jeweils ausgewählten Messniveau der optimalen Skalierung ergibt. So ergibt sich beispielsweise ein lineares Transformationsdiagramm, wenn eine Variable als numerisch behandelt wird. Als ordinal behandelte Variablen führen zu einem nicht fallenden

Transformationsdiagramm. Transformationsdiagramme für als nominal behandelte Variablen, die die Form eines “U” oder eines umgedrehten “U” aufweisen, stellen eine quadratische Beziehung dar. Nominale Variablen können auch Transformationsdiagramme ohne offensichtliche Trends ergeben, indem die Reihenfolge der Kategorien vollständig geändert wird. Folgende Abbildung zeigt ein Beispiel-Transformationsdiagramm.

Transformationsdiagramme sind besonders gut geeignet, um zu bestimmen, wie gut das ausgewählte Messniveau für die optimale Skalierung funktioniert. Wenn mehrere Kategorien ähnliche Quantifizierungen erhalten, kann die Zusammenfassung dieser Kategorien zu einer einzigen Kategorie gerechtfertigt sein. Wenn umgekehrt eine als nominal behandelte Variable Quantifizierungen erhält, die einen steigenden Trend anzeigen, kann eine ordinale Transformation zu einer ähnlichen Anpassung führen. Wenn dieser Trend linear ist, ist möglicherweise eine numerische Behandlung geeignet. Wenn jedoch die Zusammenfassung von Kategorien oder die Änderung der Messniveaus für die Skalierung angebracht ist, ändert sich die Analyse nicht in signifikanter Weise.

Abbildung 1-1
Transformationsdiagramm für Preise (numerisch)



Kategoriecodes

Bei der Kodierung kategorialer Variablen sollte mit Vorsicht vorgegangen werden, da einige Kodierungsschemata zu einer unerwünschten Datenausgabe oder zu unvollständigen Analysen führen können. Mögliche Kodierungsschemata für *Tätigkeit* werden in der folgenden Tabelle angezeigt.

Tabelle 1-2
Alternative Kodierungsschemata für die Tätigkeit

	Schema			
Kategorie	**A**	**B**	**C**	**D**
Praktikant	1	1	5	1
Vertreter	2	2	6	5
Manager	3	7	7	3

Für einige Prozeduren von Categories muss der Bereich jeder verwendeten Variablen definiert werden. Werte außerhalb dieses Bereichs werden als fehlende Werte betrachtet. Der minimale Kategoriewert ist immer 1. Der maximale Kategoriewert wird vom Benutzer angegeben. Dieser Wert gibt nicht die *Anzahl* der Kategorien für eine Variable an, sondern den *größten* Kategoriewert. In der Tabelle beispielsweise weist Schema A einen maximalen Kategoriewert von 3 und Schema B einen maximalen Kategoriewert von 7 auf, obwohl beide Schemata dieselben drei Kategorien kodieren.

Der Variablenbereich bestimmt, welche Kategorien aus der Analyse ausgeschlossen werden. Alle Kategorien mit Codes außerhalb des definierten Bereichs werden aus der Analyse ausgeschlossen. Dies ist eine einfache Methode für den Ausschluss von Kategorien, die jedoch zu unerwünschten Analysen führen kann. Wenn der maximale Kategoriewert falsch definiert ist, werden möglicherweise *gültige* Kategorien aus der Analyse ausgeschlossen. Beispiel: Bei Schema B wird durch die Festsetzung des maximalen Kategoriewerts auf den Wert 3 angegeben, dass für *Tätigkeit* Kategorien von 1 bis 3 kodiert sind. Die Kategorie *Manager* wird als fehlend behandelt. Da tatsächlich keine Kategorie als 3 kodiert wurde, enthält die dritte Kategorie in der Analyse keine Fälle. Wenn Sie alle Managerkategorien ausschließen möchten, ist diese Analyse geeignet. Wenn die Manager eingeschlossen sein sollen, muss der maximale Kategoriewert als 7 definiert sein, und fehlende Werte müssen mit Werten über 7 oder unter 1 kodiert sein.

Bei als nominal oder ordinal behandelten Variablen hat die Reihenfolge der Kategorien keine Auswirkungen auf die Ergebnisse. Bei nominalen Variablen ist nur das Label und nicht der dem Label zugeordnete Wert von Bedeutung. Bei ordinalen Variablen wird die Reihenfolge der Kategorien in den Quantifizierungen beibehalten; die Kategoriewerte selbst sind nicht von Bedeutung. Alle Kodierungsschemata, die zur selben Kategoriereihenfolge führen, weisen identische Ergebnisse auf. Beispiel: Die ersten drei Schemata in der Tabelle sind funktional äquivalent, wenn *Tätigkeit* auf ordinaler Ebene analysiert wird. Die Reihenfolge der Kategorien ist in diesen Schemata identisch. Bei Schema D dagegen werden die zweite und die dritte Kategorie invertiert, und es ergeben sich andere Ergebnisse als bei den anderen Schemata.

Obwohl viele Kodierungsschemata für eine Variable funktional äquivalent sind, werden Schemata mit kleinen Werten für die Differenz zwischen den Codes bevorzugt, da die Codes sich auf die Menge an Ausgabedaten auswirken, die sich aus der Prozedur ergeben. Alle Kategorien, die mit Werten zwischen 1 und dem benutzerdefinierten Höchstwert kodiert sind, sind gültig. Wenn eine dieser Kategorien leer ist, sind die entsprechenden Quantifizierungen je nach Prozedur entweder systemdefiniert fehlend oder 0. Obwohl sich keine dieser Zuordnungen auf die Analysen auswirkt, wird für die betreffenden Kategorien eine Ausgabe erstellt. So weist *Tätigkeit* bei Schema B vier Kategorien auf, die systemdefiniert fehlende Werte erhalten. Bei Schema C gibt es außerdem vier Kategorien, die Indikatoren für systemdefiniert fehlende Werte erhalten. Im Gegensatz dazu gibt es bei Schema A keine systemdefiniert fehlenden Quantifizierungen. Die

Verwendung aufeinander folgender Ganzzahlen als Codes für als nominal oder ordinal behandelte Variablen führt zu einer deutlich kleineren Ausgabe, ohne die Ergebnisse zu beeinträchtigen.

Die Kodierungsschemata für Variablen, die als numerisch behandelt werden, unterliegen größeren Einschränkungen, als dies bei ordinalen Variablen der Fall ist. Bei diesen Variablen ist die Differenz zwischen aufeinander folgenden Kategorien von Bedeutung. In der folgenden Tabelle werden drei Kodierungsschemata für *Alter* angezeigt.

Tabelle 1-3
Alternative Kodierungsschemata für das Alter

	Schema		
Kategorie	**A**	**B**	**C**
20	20	1	1
22	22	3	2
25	25	6	3
27	27	8	4

Bei allen Aufzeichnungen numerischer Variablen müssen die Differenzen zwischen den Kategorien erhalten bleiben. Die Verwendung der ursprünglichen Werte ist eine Methode, um die Beibehaltung der Differenzen zu gewährleisten. Dies kann jedoch dazu führen, dass viele Kategorien Indikatoren für systemdefiniert fehlende Werte aufweisen. Beispiel: Schema A verwendet die ursprünglich beobachteten Werte. Bei allen Prozeduren von Categories mit Ausnahme der Korrespondenzanalyse ist der maximale Kategoriewert 27, und der minimale Kategoriewert ist auf 1 gesetzt. Die ersten 19 Kategorien sind leer und erhalten Indikatoren für systemdefinierte fehlende Werte. Die Ausgabe kann schnell sehr umfangreich und komplex werden, wenn der maximale Kategoriewert wesentlich größer als 1 ist und sehr viele leere Kategorien zwischen 1 und dem Höchstwert liegen.

Die Größe der Datenausgabe kann durch Umkodierung verringert werden. Bei numerischen Variablen sollte die Funktion für die automatische Umkodierung jedoch nicht verwendet werden. Die Kodierung in aufeinander folgende Ganzzahlen führt zur Differenz 1 zwischen allen aufeinander folgenden Kategorien und damit dazu, dass alle Quantifizierungen dieselben Abstände aufweisen. Die metrischen Merkmale, die bei der Behandlung einer Variablen als nominal für bedeutsam erachtet werden, werden durch die Neukodierung in aufeinander folgende Ganzzahlen zerstört. Beispiel: Schema C in der Tabelle entspricht der automatischen Umkodierung von *Alter*. Die Differenz zwischen den Kategorien 22 und 25 hat sich von 3 auf 1 geändert, und die Quantifizierungen spiegeln die neue Differenz wider.

Ein alternatives Umkodierungsschema, bei dem die Differenzen zwischen den Kategorien erhalten bleiben, besteht darin, den kleinsten Kategoriewert von jeder Kategorie zu subtrahieren und zu der Differenz den Wert 1 zu addieren. Schema B ergibt sich aus dieser Transformation. Der kleinste Kategoriewert, 20, wurde von jeder Kategorie subtrahiert, und zu dem Ergebnis wurde der Wert 1 addiert. Die transformierten Codes weisen mindestens den Wert 1 auf, und alle Differenzen sind mit denen der ursprünglichen Daten identisch. Der maximale Kategoriewert ist nun 8, und die Null-Quantifizierungen vor der ersten Nicht-Null-Quantifizierung werden gelöscht. Allerdings sind die Nicht-Null-Quantifizierungen für die einzelnen Kategorien, die sich aus Schema B ergeben, mit den Quantifizierungen aus Schema A identisch.

Welche Prozedur ist für Ihre Anwendung am besten geeignet?

Die in vier dieser Prozeduren (Korrespondenzanalyse, Mehrfachkorrespondenzanalyse, kategoriale Hauptkomponentenanalyse und nichtlineare kanonische Korrelationsanalyse) enthaltenen Verfahren fallen in den allgemeinen Bereich der multivariaten Datenanalyse, die als **Dimensionsreduzierung** bekannt ist. Dies bedeutet, dass die Beziehungen zwischen den Variablen so oft wie möglich in wenigen Dimensionen – beispielsweise zwei oder drei – dargestellt werden. Dadurch können Sie Strukturen oder Muster in den Beziehungen beschreiben, die in ihrer ursprünglichen Komplexität schwer zu ergründen wären. Bei Marktforschungsanwendungen können diese Verfahren eine Art **Wahrnehmungszuordnung** sein. Ein wichtiger Vorteil dieser Prozeduren besteht darin, dass sie mit Daten mit verschiedenen Messniveaus der optimalen Skalierung umgehen können.

Die kategoriale Regression beschreibt die Beziehung zwischen einer kategorialen Responsevariablen und einer Kombination kategorialer Einflußvariablen. Der Einfluß jeder Einflußvariablen auf die Responsevariable wird durch das entsprechende Regressionsgewicht beschrieben. Wie in den anderen Prozeduren können die Daten mit verschiedenen Messniveaus der optimalen Skalierung analysiert werden.

Mit multidimensionaler Skalierung und multidimensionaler Entfaltung werden Beziehungen zwischen Objekten in einem Raum mit wenigen Dimensionen anhand der Ähnlichkeiten zwischen den Objekten beschrieben.

Es folgen kurze Richtlinien für die einzelnen Prozeduren:

- Verwenden Sie die kategoriale Regression, um die Werte einer abhängigen kategorialen Variablen aus einer Kombination unabhängiger kategorialer Variablen vorherzusagen.
- Verwenden Sie die kategoriale Hauptkomponentenanalyse, um Variationsmuster in einem einzelnen Set von Variablen mit gemischten Messniveaus der optimalen Skalierung zu berücksichtigen.
- Verwenden Sie die nichtlineare kanonische Korrelationsanalyse, um zu bewerten, wie stark zwei oder mehr Sets von Variablen mit gemischten Messniveaus der optimalen Skalierung korrelieren.
- Verwenden Sie die Korrespondenzanalyse, um zweidimensionale Kontingenztafeln oder Daten zu analysieren, die als zweidimensionale Tabelle ausgedrückt werden können, beispielsweise Daten zur bevorzugten Marke oder zur soziometrischen Wahl.
- Verwenden Sie die Mehrfachkorrespondenzanalyse zur Analyse einer Matrix mit kategorialen multivariaten Daten, wenn Sie von keiner stärkeren Annahme ausgehen möchten als von der, dass alle Variablen auf dem nominalen Niveau analysiert werden.
- Verwenden Sie die multidimensionale Skalierung zur Analyse von Ähnlichkeitsdaten, um eine Darstellung einer einzelnen Gruppe von Objekten in einem schwachdimensionierten Raum mithilfe der Methode der kleinsten Quadrate zu finden.
- Verwenden Sie die multidimensionale Entfaltung zur Analyse von Ähnlichkeitsdaten, um eine Darstellung zweier Gruppen von Objekten in einem schwachdimensionierten Raum mithilfe der Methode der kleinsten Quadrate zu finden.

Kategoriale Regression

Die Verwendung der kategorialen Regression ist am besten geeignet, wenn das Ziel der Analyse darin besteht, eine abhängige (Response-)Variable aus einem Set unabhängiger (Einfluß-)Variablen vorherzusagen. Wie bei allen Prozeduren für die optimale Skalierung werden jeder Kategorie jeder Variablen Skalenwerte zugewiesen, dergestalt, dass diese Werte in Bezug auf die Regression optimal sind. Die Lösung einer kategorialen Regression maximiert die quadrierte Korrelation zwischen der transformierten Responsevariablen und der gewichteten Kombination transformierter Einflußvariablen.

Beziehung zu anderen Prozeduren von Categories. Die kategoriale Regression mit optimaler Skalierung ist vergleichbar mit der kanonischen Korrelationsanalyse mit optimaler Skalierung und zwei Sets, von denen ein Set jeweils nur eine Variable enthält. Bei letzterem Verfahren wird die Ähnlichkeit der Sets abgeleitet, indem jedes Set mit einer unbekannten Variablen verglichen wird, die irgendwo zwischen allen Sets liegt. Bei der kategorialen Regression wird die Ähnlichkeit der transformierten Responsevariablen und der gewichteten Kombination transformierter Einflußvariablen unmittelbar bewertet.

Beziehung zu Standardverfahren. Beim Standardverfahren der linearen Regression können kategoriale Variablen entweder als Indikatorvariablen umkodiert oder genauso wie Variablen auf Intervallniveau behandelt werden. Beim ersten Ansatz enthält das Modell eine gesonderte Konstante und Steigung für die einzelnen Niveaukombinationen für die kategorialen Variablen. Dies führt zu einer großen Anzahl zu interpretierender Parameter. Im zweiten Ansatz wird für jede Variable nur ein einziger Parameter geschätzt. Durch die willkürliche Festlegung der Kategoriekodierungen werden jedoch Verallgemeinerungen unmöglich.

Wenn einige Variablen nicht stetig sind, stehen alternative Analysen zur Verfügung. Bei stetiger Responsevariabler und kategorialen Einflußvariablen wird häufig eine Varianzanalyse durchgeführt. Bei kategorialer Responsevariabler und stetigen Einflußvariablen kann eine logistische Regression oder eine Diskriminanzanalyse angemessen sein. Wenn sowohl die Responsevariable als auch die Einflußvariablen kategorial sind, werden häufig loglineare Modelle verwendet.

Die Regression mit optimaler Skalierung bietet drei Messniveaus für die Skalierung der einzelnen Variablen. Durch Kombinationen dieser Niveaus kann eine große Bandbreite an nichtlinearen Beziehungen behandelt werden, für die eine einzelne “Standard”-Methode nicht geeignet ist. Folglich bietet die optimale Skalierung eine größere Flexibilität als die Standardansätze bei kaum höherer Komplexität.

Außerdem werden durch nichtlineare Transformationen der Einflußvariablen normalerweise die Abhängigkeiten zwischen den einzelnen Einflußvariablen verringert. Wenn Sie die Eigenwerte der Korrelationsmatrix für die Einflußvariablen mit den Eigenwerten der Korrelationsmatrix für die optimal skalierten Einflußvariablen vergleichen, ist die zweite Gruppe zumeist weniger variabel als die erste. Mit anderen Worten: bei der kategorialen Regression werden durch die optimale Skalierung die großen Eigenwerte der Korrelationsmatrix für die Einflußvariablen kleiner und die kleinen Eigenwerte größer.

Kategoriale Hauptkomponentenanalyse

Die kategoriale Hauptkomponentenanalyse ist am besten geeignet, wenn Variationsmuster in einem einzelnen Set von Variablen mit gemischten Messniveaus der optimalen Skalierung berücksichtigt werden sollen. Dieses Verfahren versucht, die Dimensionalität eines Variablen-Sets zu verringern und dabei gleichzeitig die Variation soweit wie möglich zu berücksichtigen. Jeder Kategorie jeder Variablen werden Skalenwerte zugewiesen, dergestalt, dass diese Werte in Bezug auf die Hauptkomponentenlösung optimal sind. Die Objekte in der Analyse erhalten Komponentenwerte auf der Grundlage der quantifizierten Daten. Diagramme der Komponentenwerte lassen Muster in Bezug auf die Objekte in der Analyse sichtbar werden und können ungewöhnliche Objekte in den Daten kenntlich machen. Die Lösung einer kategorialen Hauptkomponentenanalyse maximiert die Korrelationen der Objektwerte mit den einzelnen quantifizierten Variablen für die Anzahl der angegebenen Komponenten (Dimensionen).

Eine wichtige Anwendung der kategorialen Hauptkomponentenanalyse ist die Untersuchung von Präferenzdaten, bei der die Befragten eine Reihe von Items nach Ihrer Präferenz einstufen oder bewerten. Bei der üblichen SPSS-Datenkonfiguration stellen die Zeilen Einzelwerte und die Spalten Maße für die Items dar. Die zeilenübergreifenden Scores sind Präferenzscores (beispielsweise auf einer Skala von 0 bis 10). Dadurch werden die Daten zeilenabhängig. Bei Präferenzdaten kann es sinnvoll sein, die Einzelwerte als Variablen zu behandeln. Mit der Prozedur "Transponieren" können Sie die Daten transponieren. Die Bewerter werden die Variablen, und alle Variablen werden als ordinal deklariert. In CATPCA können problemlos mehr Variablen als Objekte verwendet werden.

Beziehung zu anderen Prozeduren von Categories. Wenn alle Variablen als mehrfach nominal deklariert werden, führt die kategoriale Hauptkomponentenanalyse zu einer Analyse, die einer Mehrfachkorrespondenzanalyse, die für dieselben Variablen durchgeführt wurde, entspricht. Daher kann die kategoriale Hauptkomponentenanalyse als eine Art Mehrfachkorrespondenzanalyse betrachtet werden, bei der einige der Variablen als ordinal oder numerisch deklariert sind.

Beziehung zu Standardverfahren. Wenn alle Variablen auf dem numerischen Niveau skaliert sind, entspricht die kategoriale Hauptkomponentenanalyse der normalen Hauptkomponentenanalyse.

Allgemeiner gesagt: Die kategoriale Hauptkomponentenanalyse ist eine Alternative zur Berechnung der Korrelationen zwischen nichtnumerischen Skalen und ihrer Analyse mithilfe der normalen Hauptkomponentenanalyse oder der Faktoranalyse. Die unreflektierte Verwendung des üblichen Korrelationskoeffizienten nach Pearson als Zusammenhangsmaß für ordinale Daten kann zu einer nicht trivialen Verzerrung in der Schätzung der Korrelationen führen.

Nichtlineare kanonische Korrelationsanalyse

Die nichtlineare kanonische Korrelationsanalyse ist eine sehr allgemeine Prozedur mit vielen verschiedenen Anwendungen. Ziel der nichtlinearen kanonischen Korrelationsanalyse ist die Analyse der Beziehungen zwischen zwei oder mehr Sets von Variablen, anstatt zwischen den Variablen selbst (wie bei der Hauptkomponentenanalyse). So können beispielsweise zwei Variablen-Sets vorliegen, wobei es sich bei einem Variablen-Set um Items zum demografischen Hintergrund für eine Gruppe von Befragten handelt, während das zweite Set aus Antworten auf eine Gruppe von Items zur Einstellung der Personen besteht. Die Skalierungsniveaus in der Analyse können aus einer beliebigen Mischung aus nominalen, ordinalen und numerischen

Messniveaus bestehen. Die kanonische Korrelationsanalyse mit optimaler Skalierung bestimmt die Ähnlichkeit zwischen den Sets und vergleicht gleichzeitig die kanonischen Variablen aus den einzelnen Sets mit einem Kompromiß-Set, das aus den Werten besteht, die den Objekten zugewiesen wurden.

Beziehung zu anderen Prozeduren von Categories. Wenn zwei oder mehr Variablen-Sets mit jeweils nur einer einzigen Variablen vorliegen, entspricht die kanonische Korrelationsanalyse mit optimaler Skalierung der Hauptkomponentenanalyse mit optimaler Skalierung. Wenn alle Variablen in einer Analyse mit einer Variablen pro Set mehrfach nominal sind, entspricht die kanonische Korrelationsanalyse mit optimaler Skalierung der Mehrfachkorrespondenzanalyse. Bei zwei Variablen-Sets, von denen eine nur eine einzige Variable enthält, entspricht die kanonische Korrelationsanalyse mit optimaler Skalierung der kategorialen Regression mit optimaler Skalierung.

Beziehung zu Standardverfahren. Die kanonische Standard-Korrelationsanalyse ist ein statistisches Verfahren, bei der eine lineare Kombination eines Variablen-Sets und eine lineare Kombination eines zweiten Variablen-Sets ermittelt werden, die maximal korreliert sind. Ausgehend von diesem Set mit linearen Kombinationen kann die kanonische Korrelationsanalyse darauf folgende unabhängige Sets mit linearen Kombinationen ermitteln (als kanonische Variablen bezeichnet), bis zu einer Maximalzahl, die der Anzahl der Variablen im kleineren Set entspricht.

Wenn in der Analyse zwei Variablen-Sets vorliegen und alle Variablen als numerisch definiert sind, entspricht die kanonische Korrelationsanalyse mit optimaler Skalierung einer kanonischen Standard-Korrelationsanalyse. SPSS beinhaltet zwar keine Prozedur für die kanonische Korrelationsanalyse, doch viele der relevanten Statistiken können aus einer multivariaten Varianzanalyse gewonnen werden.

Die kanonische Korrelationsanalyse mit optimaler Skalierung bietet noch viele weitere Anwendungsmöglichkeiten. Bei zwei Variablen-Sets, von denen eines eine nominale Variable enthält, die als einfach nominal deklariert ist, können die Ergebnisse der kanonischen Korrelationsanalyse mit optimaler Skalierung ähnlich wie die Ergebnisse einer Regressionsanalyse interpretiert werden. Wenn Sie die Variable als mehrfach nominal betrachten, ist die Analyse mit optimaler Skalierung eine Alternative zur Diskriminanzanalyse. Durch die Gruppierung der Variablen in mehr als zwei Sets erhalten Sie eine Vielzahl von Möglichkeiten zur Analyse der Daten.

Korrespondenzanalyse

Ziel der Korrespondenzanalyse ist die Erstellung von Biplots für Korrespondenztafeln. In einer Korrespondenztafel wird davon ausgegangen, dass die Zeilen- und Spaltenvariablen ungeordnete Kategorien darstellen; daher wird hier immer das nominale Messniveau der optimalen Skalierung verwendet. Beide Variablen werden hinsichtlich ihrer nominalen Informationen untersucht. Es wird also nur die Tatsache berücksichtigt, dass sich einige Objekte in derselben Kategorie befinden, andere jedoch nicht. Über die Distanz oder die Reihenfolge der verschiedenen Kategorien derselben Variablen werden keine Annahmen gemacht.

Eine spezielle Verwendung der Korrespondenzanalyse besteht in der Analyse von zweidimensionalen Kontingenztafeln. Wenn eine Tabelle *r* aktive Zeilen und *c* aktive Spalten aufweist, beträgt die Anzahl der Dimensionen in der Korrespondenzanalyse mindestens d *r* minus 1 bzw. *c* minus 1 (der kleinere der beiden Werte wird verwendet). Anders

ausgedrückt: Die Zeilen- oder Spaltenkategorien einer Kontingenztafel könnten perfekt in einem mehrdimensionalen Raum dargestellt werden. In der Praxis würde man jedoch die Zeilen- und Spaltenkategorien einer zweidimensionalen Tabelle in einem Raum mit wenigen, beispielsweise zwei, Dimensionen darstellen wollen, da zweidimensionale Diagramme leichter verständlich sind als mehrdimensionale räumliche Darstellungen.

Wenn weniger als die Höchstzahl der möglichen Dimensionen verwendet werden, beschreibt die in der Analyse erstellte Statistik, wie gut die Zeilen- und Spaltenkategorien in der flachdimensionierten Darstellung repräsentiert werden. Vorausgesetzt, die Darstellung der zweidimensionalen Lösung weist eine gute Qualität auf, können Sie Plots der Zeilenpunkte und der Spaltenpunkte untersuchen, um herauszufinden, welche Kategorien der Zeilenvariablen ähnlich sind, welche Kategorien der Spaltenvariablen ähnlich sind und welche Zeilen- und Spaltenkategorien einander ähnlich sind.

Beziehung zu anderen Prozeduren von Categories. Die einfache Korrespondenzanalyse ist auf zweidimensionale Tabellen beschränkt. Liegen mehr als zwei relevante Variablen vor, können Sie die Variablen zu Interaktionsvariablen zusammenfassen. Bei den Variablen *Region*, *Tätigkeit* und *Alter* beispielsweise könnten *Region* und *Tätigkeit* zu der neuen Variablen *Regtätigkeit* zusammengefaßt werden, die, wie in der untenstehenden Tabelle gezeigt, 12 Kategorien aufweist. Diese neue Variable führt zu einer zweidimensionalen Tabelle mit *Alter* (12 Zeilen, 4 Spalten), die mithilfe der Korrespondenzanalyse analysiert werden kann.

Tabelle 1-4
Kombinationen von Region und Tätigkeit

Kategoriecode	Kategoriedefinition	Kategoriecode	Kategoriedefinition
1	Norden, Praktikant	7	Osten, Praktikant
2	Norden, Vertreter	8	Osten, Vertreter
3	Norden, Manager	9	Osten, Manager
4	Süden, Praktikant	10	Westen, Praktikant
5	Süden, Vertreter	11	Westen, Vertreter
6	Süden, Manager	12	Westen, Manager

Ein Nachteil bei diesem Ansatz besteht darin, dass beliebige Variablenpaare zusammengefaßt werden können. *Tätigkeit* und *Alter* können kombiniert werden, was ebenfalls eine Variable mit 12 Kategorien ergibt. Oder man kann *Region* und *Alter* kombinieren, was zu einer neuen Variablen mit 16 Kategorien führt. Alle diese Interaktionsvariablen ergeben zusammen mit der verbleibenden Variablen eine zweidimensionale Tabelle. Die Korrespondenzanalysen dieser drei Tabellen führen nicht zu identischen Ergebnissen, aber dennoch ist jeder der Ansätze gültig. Außerdem können bei vier oder mehr Variablen zweidimensionale Tabellen erstellt werden, die eine Interaktionsvariable mit einer weiteren Interaktionsvariablen vergleichen. Die Anzahl der möglichen Tabellen für die Analyse kann recht groß werden, selbst bei nur relativ wenigen Variablen. Sie können eine dieser Tabellen für die Analyse auswählen oder alle analysieren. Alternativ kann die Prozedur der Mehrfachkorrespondenzanalyse verwendet werden, um alle Variablen gleichzeitig zu untersuchen, ohne dass Interaktionsvariablen konstruiert werden müssen.

Beziehung zu Standardverfahren. Mit der Prozedur Crosstabs können ebenfalls Kontingenztabellen analysiert werden, wobei Unabhängigkeit einen gemeinsamen Fokus der Analysen darstellt. Selbst bei kleinen Tabellen kann es jedoch schwierig sein, die Ursache von Abweichungen von

der Unabhängigkeit zu ermitteln. Der Nutzen der Korrespondenzanalyse besteht in der Anzeige derartiger Muster für zweidimensionale Tabellen beliebiger Größe. Wenn ein Zusammenhang zwischen den Zeilen- und den Spaltenvariablen besteht, d. h., wenn der Chi-Quadrat-Wert signifikant ist, kann die Korrespondenzanalyse die Ermittlung der Art dieser Beziehung erleichtern.

Mehrfachkorrespondenzanalyse

Bei der Mehrfachkorrespondenzanalyse wird versucht, eine Lösung zu erstellen, bei der Objekte innerhalb derselben Kategorie eng zusammen und Objekte in verschiedenen Kategorien in großem Abstand zueinander geplottet werden. Jedes Objekt liegt so nah wie möglich an den Kategoriepunkten der Kategorien, die für dieses Objekt gelten. So unterteilen die Kategorien die Objekte in homogene Untergruppen. Variablen werden als homogen betrachtet, wenn sie Objekte, die in denselben Kategorien vorliegen, auch in dieselben Untergruppen einordnen.

Bei einer eindimensionalen Lösung weist die Mehrfachkorrespondenzanalyse jeder Kategorie jeder Variablen optimale Skalenwerte (Kategoriequantifikationen) zu, dergestalt, dass die Kategorien insgesamt im Durchschnitt die maximale Streubreite aufweisen. Bei einer zweidimensionalen Lösung ermittelt die Mehrfachkorrespondenzanalyse ein zweites Set von Quantifizierungen der Kategorien jeder Variablen, das nicht mit dem ersten Set in Zusammenhang steht, wobei wieder versucht wird, die Streubreite zu maximieren, usw. Da die Kategorien einer Variablen so viele Bewertungen erhalten, wie Dimensionen vorhanden sind, wird davon ausgegangen, dass die Variablen in der Analyse auf dem Messniveau der optimalen Skalierung mehrfach nominal sind.

Die Mehrfachkorrespondenzanalyse weist auch den Objekten in der Analyse solche Werte zu, dass die Kategoriequantifizierungen die Durchschnittswerte bzw. Zentroide der Objektwerte der Objekte in dieser Kategorie sind.

Beziehung zu anderen Prozeduren von Categories. Die Mehrfachkorrespondenzanalyse wird auch als Homogenitätsanalyse oder Doppelskalierung bezeichnet. Sie führt zu vergleichbaren (jedoch nicht identischen) Lösungen wie die Korrespondenzanalyse, wenn nur zwei Variablen vorhanden sind. Die Korrespondenzanalyse führt zu einer eindeutigen Ausgabe, bei der die Anpassung und die Qualität der Darstellung der Lösung, einschließlich der Stabilitätsinformationen, zusammengefaßt werden. Daher ist bei zwei Variablen die Korrespondenzanalyse normalerweise der Mehrfachkorrespondenzanalyse vorzuziehen. Ein weiterer Unterschied zwischen den beiden Prozeduren besteht darin, dass es sich bei der Eingabe für die Mehrfachkorrespondenzanalyse um eine Datenmatrix handelt, bei der die Zeilen Objekte und die Spalten Variablen sind, wohingegen es sich bei der Eingabe für die Korrespondenzanalyse um dieselbe Datenmatrix, eine allgemeine Distanzmatrix oder eine verbundene Kontingenztafel handeln kann, die eine aggregierte Matrix ist, bei der sowohl die Zeilen als auch die Spalten Kategorien von Variablen repräsentieren. Die Mehrfachkorrespondenzanalyse kann auch als Hauptkomponentenanalyse von Daten betrachtet werden, die auf dem mehrfach nominalen Niveau skaliert wurden.

Beziehung zu Standardverfahren. Die Mehrfachkorrespondenzanalyse kann als Analyse einer mehrdimensionalen Kontingenztafel betrachtet werden. Mehrdimensionale Kontingenztafeln können auch mit der Prozedur Crosstabs analysiert werden, doch Crosstabs gibt getrennte Zusammenfassungsstatistiken für die einzelnen Kategorien jeder Kontrollvariablen aus. Bei der Mehrfachkorrespondenzanalyse kann häufig die Beziehung zwischen allen Variablen in einem

einzelnen zweidimensionalen Diagramm zusammengefaßt werden. Eine erweiterte Verwendung der Mehrfachkorrespondenzanalyse besteht darin, die ursprünglichen Kategoriewerte mit den optimalen Skalenwerten aus der ersten Dimension zu ersetzen und eine sekundäre multivariate Analyse durchzuführen. Da die Mehrfachkorrespondenzanalyse die Kategoriebeschriftungen durch numerische Skalenwerte ersetzt, können nach der Mehrfachkorrespondenzanalyse viele verschiedene Prozeduren angewendet werden, für die numerische Daten erforderlich sind. Die Prozedur "Faktorenanalyse" beispielsweise erstellt zunächst eine erste Hauptkomponente, die der ersten Dimension der Mehrfachkorrespondenzanalyse entspricht. Die Komponentenwerte in der ersten Dimension sind gleich den Objektwerten, und die quadrierten Komponentenladungen sind gleich den Diskriminationsmaßen. Die zweite Dimension der Mehrfachkorrespondenzanalyse jedoch ist nicht gleich der zweiten Dimension der Faktorenanalyse.

Multidimensionale Skalierung

Die multidimensionale Skalierung ist am besten geeignet, wenn das Ziel der Analyse darin besteht, die Struktur eines Sets von Distanzmaßen innerhalb einer einzelnen Gruppe von Objekten oder Fällen zu ermitteln. Dies wird durch das Zuweisen von Beobachtungen zu bestimmten Positionen in einem konzeptuellen Raum mit wenigen Dimensionen erzielt, wobei die Abstände zwischen den Punkten des Raumes mit den vorgegebenen Unähnlichkeiten bzw. Ähnlichkeiten so gut wie möglich übereinstimmen. Als Ergebnis werden die Objekte in diesem Raum mit wenigen Dimensionen mithilfe der Methode der kleinsten Quadrate dargestellt, was häufig zu einem besseren Verständnis der Daten beiträgt.

Beziehung zu anderen Prozeduren von Categories. Wenn multivariate Daten vorliegen, aus denen Sie Distanzen erstellen und anschließend mit multidimensionaler Skalierung analysieren, erhalten Sie ein ähnliches Ergebnis wie bei der Analyse der Daten mithilfe der kategorialen Hauptkomponentenanalyse mit Objektprinzipal-Normalisierung. Diese Art von PCA ist auch als Hauptkoordinatenanalyse bekannt.

Beziehung zu Standardverfahren. Die Prozedur "Multidimensionale Skalierung" von Categories (PROXSCAL) bietet zahlreiche Verbesserungen gegenüber der in SPSS Base verfügbaren Skalierungsprozedur (ALSCAL). PROXSCAL bietet einen schnelleren Algorithmus für bestimmte Modelle und ermöglicht Ihnen, Einschränkungen für den gemeinsamen Raum festzulegen. Außerdem versucht PROXSCAL, den normalisierten Roh-Stress zu minimieren und nicht den S-Stress (auch als **Belastung** bezeichnet). Der normalisierte Roh-Stress wird im allgemeinen bevorzugt, da es sich dabei um ein Maß handelt, das auf den Distanzen beruht, während S-Stress auf den quadrierten Distanzen beruht.

Multidimensionale Entfaltung

Die multidimensionale Entfaltung ist am besten geeignet, wenn das Ziel der Analyse darin besteht, die Struktur eines Sets von Distanzmaßen zwischen zwei Gruppen von Objekten (den Zeilen- und Spaltenobjekten) zu ermitteln. Dies wird durch das Zuweisen von Beobachtungen zu bestimmten Positionen in einem konzeptuellen Raum mit wenigen Dimensionen erzielt, wobei die Abstände zwischen den Punkten des Raumes mit den vorgegebenen Unähnlichkeiten bzw. Ähnlichkeiten so gut wie möglich übereinstimmen. Als Ergebnis werden die Zeilen- und Spaltenobjekte in

diesem schwachdimensionierten Raum mithilfe der Methode der kleinsten Quadrate dargestellt, was häufig zu einem besseren Verständnis der Daten beiträgt.

Beziehung zu anderen Prozeduren von Categories. Wenn die Daten aus Distanzen innerhalb einer einzelnen Gruppe von Objekten (eine quadratische und symmetrische Matrix) bestehen, verwenden Sie die multidimensionale Skalierung.

Beziehung zu Standardverfahren. Die Prozedur "Multidimensionale Entfaltung" in SPSS Categories (PREFSCAL) bietet zahlreiche Verbesserungen gegenüber der in SPSS Base verfügbaren Entfaltungsprozedur (über ALSCAL). Mit PREFSCAL kann der gemeinsame Raum eingeschränkt werden. Außerdem wird in PREFSCAL versucht, ein penalisiertes Stressmaß zu minimieren, was dazu beiträgt, degenerierte Lösungen zu vermeiden (für die ältere Algorithmen anfällig sind).

Seitenverhältnis in Diagrammen mit optimaler Skalierung

Das Seitenverhältnis in Diagrammen mit optimaler Skalierung ist isotropisch. Bei einem zweidimensionalen Diagramm ist die Distanz, die in Dimension 1 für eine (1) Einheit steht, gleich der Distanz, die in Dimension 2 für eine (1) Einheit steht. Wenn Sie bei einem zweidimensionalen Diagramm den Bereich einer Dimension ändern, ändert das System die Größe der anderen Dimension, um die physischen Abstände gleich groß zu halten. Das isotropische Seitenverhältnis kann für die Prozeduren für die optimale Skalierung nicht außer Kraft gesetzt werden.

Empfohlene Literatur

In folgenden Texten finden Sie allgemeine Informationen zu den Verfahren zur optimalen Skalierung:

Barlow, R. E., D. J. Bartholomew, D. J. Bremner, als auch H. D. Brunk. 1972. *Statistical inference under order restrictions*. New York: John Wiley and Sons.

Benzécri, J. P. 1969. Statistical analysis as a tool to make patterns emerge from data. In: *Methodologies of Pattern Recognition,* S. Watanabe (Hg.). New York: Academic Press, 35–74.

Bishop, Y. M., S. E. Feinberg, als auch P. W. Holland. 1975. *Discrete multivariate analysis: Theory and practice*. Cambridge, Massachusetts: MIT Press.

De Leeuw, J. 1984. The Gifi system of nonlinear multivariate analysis. In: *Data Analysis and Informatics III,* E. Diday, et al. (Hg.)., 415–424.

De Leeuw, J. 1990. Multivariate analysis with optimal scaling. In: *Progress in Multivariate Analysis,* S. Das Gupta, als auch J. Sethuraman (Hgg.). Kalkutta: Indian Statistical Institute.

De Leeuw, J., als auch J. Van Rijckevorsel. 1980. HOMALS and PRINCALS—Some generalizations of principal components analysis. In: *Data Analysis and Informatics,* E. Diday, et al. (Hg.). Amsterdam: North-Holland, 231–242.

De Leeuw, J., F. W. Young, als auch Y. Takane. 1976. Additive structure in qualitative data: An alternating least squares method with optimal scaling features. *Psychometrika*, 41, 471–503.

Gifi, A. 1990. *Nonlinear multivariate analysis*. Chichester: John Wiley and Sons.

Heiser, W. J., als auch J. J. Meulman. 1995. Nonlinear methods for the analysis of homogeneity and heterogeneity. In: *Recent Advances in Descriptive Multivariate Analysis,* W. J. Krzanowski (Hg.). Oxford: Oxford University Press, 51–89.

Israëls, A. 1987. *Eigenvalue techniques for qualitative data*. Leiden: DSWO Press.

Krzanowski, W. J., als auch F. H. C. Marriott. 1994. *Multivariate analysis: Part I, distributions, ordination and inference*. London: Edward Arnold.

Lebart, L., A. Morineau, als auch K. M. Warwick. 1984. *Multivariate descriptive statistical analysis*. New York: John Wiley and Sons.

Max, J. 1960. Quantizing for minimum distortion. *Proceedings IEEE (Information Theory)*, 6, 7–12.

Meulman, J. J. 1986. *A distance approach to nonlinear multivariate analysis*. Leiden: DSWO Press.

Meulman, J. J. 1992. The integration of multidimensional scaling and multivariate analysis with optimal transformations of the variables. *Psychometrika*, 57, 539–565.

Nishisato, S. 1980. *Analysis of categorical data: Dual scaling and its applications*. Toronto: University of Toronto Press.

Nishisato, S. 1994. *Elements of dual scaling: An introduction to practical data analysis*. Hillsdale, N.J.: Lawrence Erlbaum Associates, Inc.

Rao, C. R. 1973. *Linear statistical inference and its applications*, 2 (Hg.). New York: John Wiley and Sons.

Rao, C. R. 1980. Matrix approximations and reduction of dimensionality in multivariate statistical analysis. In: *Multivariate Analysis, Vol. 5,* P. R. Krishnaiah (Hg.). Amsterdam: North-Holland, 3–22.

Roskam, E. E. 1968. *Metric analysis of ordinal data in psychology*. Voorschoten: VAM.

Shepard, R. N. 1966. Metric structures in ordinal data. *Journal of Mathematical Psychology*, 3, 287–315.

Wolter, K. M. 1985. *Introduction to variance estimation*. Berlin: Springer-Verlag.

Young, F. W. 1981. Quantitative analysis of qualitative data. *Psychometrika*, 46, 357–387.

Kapitel
2

Kategoriale Regression (CATREG)

Durch die **Kategoriale Regression** werden kategoriale Daten quantifiziert, indem den Kategorien numerische Werte zugewiesen werden. Dadurch ergibt sich für die transformierten Variablen eine optimale lineare Regressionsgleichung. Die kategoriale Regression wird auch mit dem Akronym CATREG (*Cat*egorical *Reg*ression) bezeichnet.

Das Standardverfahren der linearen Regressionsanalyse beinhaltet die Minimierung der Summe von quadrierten Differenzen zwischen einer Antwortvariablen (abhängig) und einer gewichteten Kombination von Einflußvariablen (unabhängig). Variablen sind in der Regel quantitativ, wobei (nominale) kategoriale Daten in Binär- oder Kontrastvariablen umkodiert werden. Infolgedessen dienen kategoriale Variablen einer Aufteilung in verschiedene Gruppen von Fällen, sodass jeweils separate Parametersätze für jede Gruppe geschätzt werden. Die geschätzten Koeffizienten geben die Auswirkung einer Änderung in den Einflußvariablen auf die Antwortvariable wieder. Die Antwort kann für jede beliebige Kombination von Einflußwerten vorhergesagt werden.

Eine andere Methode besteht darin, dass die Antwort auf die kategorialen Einflußwerte selbst einer Regression unterzogen wird. Folglich wird für jede Variable ein Koeffizient geschätzt. Bei kategorialen Variablen sind die Kategoriewerte jedoch willkürlich. Durch verschiedene Kodierungsarten der Kategorien ergeben sich jeweils unterschiedliche Koeffizienten, wodurch ein analysenübergreifender Vergleich gleicher Variablen erschwert wird.

CATREG erweitert die Standardmethode durch eine gleichzeitige Skalierung nominaler, ordinaler und numerischer Variablen. Die Prozedur quantifiziert kategoriale Variablen, sodass in den Quantifikationen die Merkmale der ursprünglichen Kategorien zum Ausdruck kommen. Dadurch werden quantifizierte kategoriale Variablen auf dieselbe Weise wie numerische Variablen behandelt. Durch die Verwendung nichtlinearer Transformationen können Variablen auf einer Vielzahl von Ebenen analysiert und somit das jeweils geeignetste Modell gefunden werden.

Beispiel. Mithilfe der kategorialen Regression könnte beispielsweise beschrieben werden, wie die Zufriedenheit im Beruf von der Berufsgruppe, der geografischen Region und der Fahrtzeit abhängt. Unter Umständen könnte sich herausstellen, dass bei leitenden Angestellten und geringer Fahrtzeit zum Arbeitsplatz ein hohes Zufriedenheitsniveau zu verzeichnen ist. Mithilfe der resultierenden Regressionsgleichung könnte die Zufriedenheit im Beruf für eine beliebige Kombination der drei unabhängigen Variablen prognostiziert werden.

Statistiken und Diagramme. Häufigkeiten, Regressionskoeffizienten, ANOVA-Tabelle, Iterationsprotokoll, Kategorienquantifikationen, Korrelationen zwischen nicht transformierten Einflußvariablen, Korrelationen zwischen transformierten Einflußvariablen, Residuen-Diagramme und Transformationsdiagramme.

Daten. Mit CATREG werden Kategorienummervariablen verarbeitet. Diese Kategorienummern müssen positive Ganzzahlen sein. Im Dialogfeld “Diskretisierung” können Sie Variablen mit Dezimalbrüchen und String-Variablen in positive Ganzzahlen umwandeln.

Annahmen. Es ist nur eine Antwortvariable zulässig, die maximale Anzahl der Einflussvariablen beträgt jedoch 200. Die Daten müssen mindestens drei gültige Fälle enthalten. Die Anzahl gültiger Fälle muss die Anzahl der Einflussvariablen plus eins übersteigen.

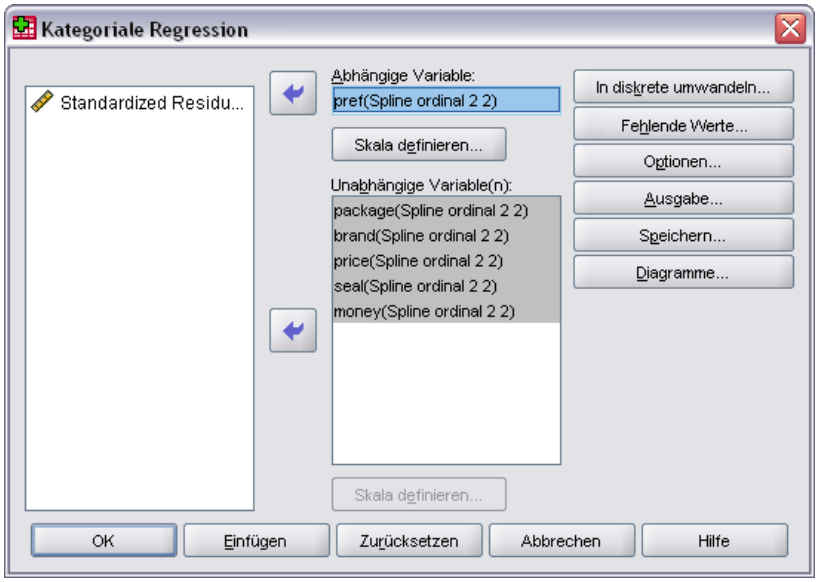
Verwandte Prozeduren. CATREG entspricht der kategorialen kanonischen Korrelationsanalyse mit optimaler Skalierung (OVERALS) und zwei Sets, von denen ein Set jeweils nur eine Variable enthält. Eine Skalierung aller Variablen auf numerischem Niveau entspricht der normalen Mehrfachregressionsanalyse.

So lassen Sie eine kategoriale Regression berechnen:

- Wählen Sie die folgenden Befehle aus den Menüs aus:

Analysieren
  Regression
    Optimale Skalierung...

Abbildung 2-1
Dialogfeld "Kategoriale Regression"



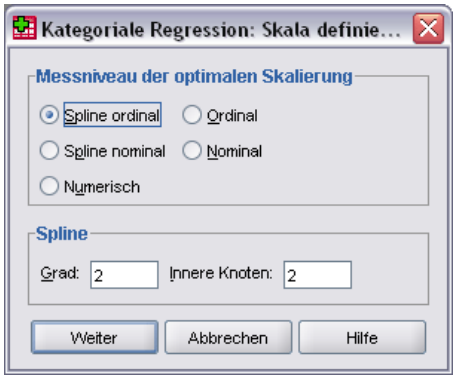
- Wählen Sie die abhängige Variable und die unabhängige(n) Variable(n).
- Klicken Sie auf OK.

Ändern Sie bei Bedarf das Skalierungsniveau für die einzelnen Variablen.

Definieren der Skala in der kategorialen Regression

Sie können das Messniveau der optimalen Skalierung für die abhängigen und unabhängigen Variablen festlegen. In der Standardeinstellung werden sie als monotone Splines zweiten Grades (ordinal) mit zwei inneren Knoten skaliert. Außerdem können Sie die Gewichtung der Analysevariablen bestimmen.

Abbildung 2-2
Dialogfeld "Skala definieren"



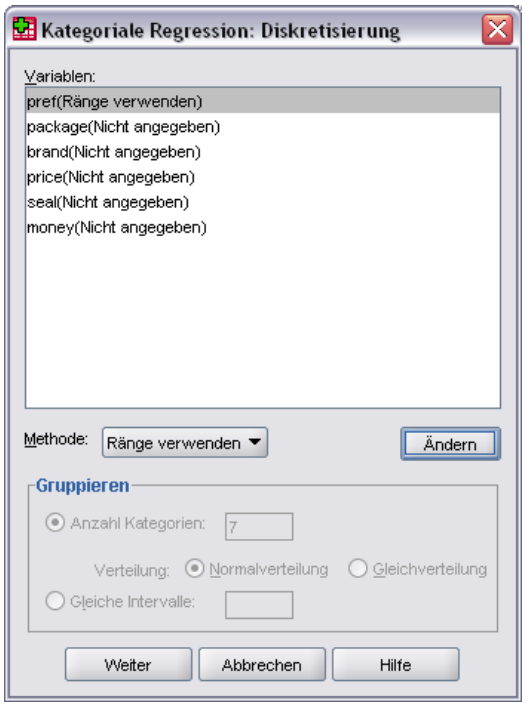
Messniveau der optimalen Skalierung. Sie können auch das Messniveau zur Quantifizierung jeder Variablen auswählen.

- **Spline ordinal.** Die Reihenfolge der Kategorien bei der beobachteten Variable wird bei der optimal skalierten Variable beibehalten. Die Kategorienpunkte werden auf einer geraden Linie (einem Vektor) durch den Koordinatenursprung plaziert. Die resultierende Transformation ist ein glatter, stückweiser, monotoner Polynom mit dem gewählten Grad. Die Stücke werden durch die vom Benutzer festgelegte Anzahl und die durch die Prozedur bestimmte Lage der inneren Knoten definiert.
- **Spline nominal.** Die einzige Information aus der beobachteten Variable, die in der optimal skalierten Variable beibehalten wird, ist die Gruppierung der Objekte in Kategorien. Die Reihenfolge der Kategorien bei der beobachteten Variablenwird nicht beibehalten. Die Kategorienpunkte werden auf einer geraden Linie (einem Vektor) durch den Koordinatenursprung plaziert. Die resultierende Transformation ist ein glatter, stückweiser, möglicherweise nichtmonotoner Polynom mit dem gewählten Grad. Die Stücke werden durch die vom Benutzer festgelegte Anzahl und die durch die Prozedur bestimmte Lage der inneren Knoten definiert.
- **Ordinal.** Die Reihenfolge der Kategorien bei der beobachteten Variable wird bei der optimal skalierten Variable beibehalten. Die Kategorienpunkte werden auf einer geraden Linie (einem Vektor) durch den Koordinatenursprung plaziert. Die resultierende Transformation ist besser angepaßt, aber weniger glatt als bei "Spline ordinal".
- **Nominal.** Die einzige Information aus der beobachteten Variable, die in der optimal skalierten Variable beibehalten wird, ist die Gruppierung der Objekte in Kategorien. Die Reihenfolge der Kategorien bei der beobachteten Variablenwird nicht beibehalten. Die Kategorienpunkte werden auf einer geraden Linie (einem Vektor) durch den Koordinatenursprung plaziert. Die resultierende Transformation ist besser angepaßt, aber weniger glatt als bei "Spline nominal".
- **Numerisch.** Die Kategorien werden als geordnet mit gleichen Abständen behandelt (Intervallniveau). Die Reihenfolge der Kategorien und die gleichen Abstände zwischen den Kategorienummern in der beobachteten Variablenbleiben in der optimal skalierten Variablen erhalten. Die Kategorienpunkte werden auf einer geraden Linie (einem Vektor) durch den Koordinatenursprung plaziert. Wenn alle Variablen auf numerischem Niveau skaliert sind, entspricht die Analyse der normalen Hauptkomponentenanalyse.

Kategoriale Regression: Diskretisierung

Im Dialogfeld “Kategoriale Hauptkomponenten: Diskretisierung” können Sie eine Methode zum Umkodieren der Variablen festlegen. Sofern nicht anders angegeben, werden Variablen mit Dezimalwerten in sieben Kategorien gruppiert, die nahezu normalverteilt sind. (Wenn die Variable weniger als sieben unterschiedliche Werte aufweist, entspricht die Anzahl der Kategorien der Anzahl der unterschiedlichen Werte.) Werte von String-Variablen werden immer in positive ganze Zahlen umgewandelt, indem in aufsteigender alphanumerischer Reihenfolge Kategorienummern zugewiesen werden. Die Diskretisierung von String-Variablen erfolgt anhand dieser Ganzzahlen. Andere Variablen werden in der Standardeinstellung nicht modifiziert. Die diskretisierten Variablen werden dann in der Analyse verwendet.

Abbildung 2-3
Dialogfeld “Diskretisierung”



Methode. Verfügbar sind die Optionen “Gruppieren”, “Einstufen” oder “Multiplizieren”.

- **Gruppieren.** Hierbei werden die Daten in eine bestimmte Anzahl von Kategorien oder nach Intervall umkodiert.
- **Einstufen.** Die Variable wird durch Bilden einer Rangfolge diskretisiert.
- **Multiplizieren.** Die aktuellen Werte der Variablen werden standardisiert, mit 10 multipliziert und gerundet. Anschließend wird eine Konstante zu den Werten addiert, sodass der niedrigste diskretisierte Wert 1 beträgt.

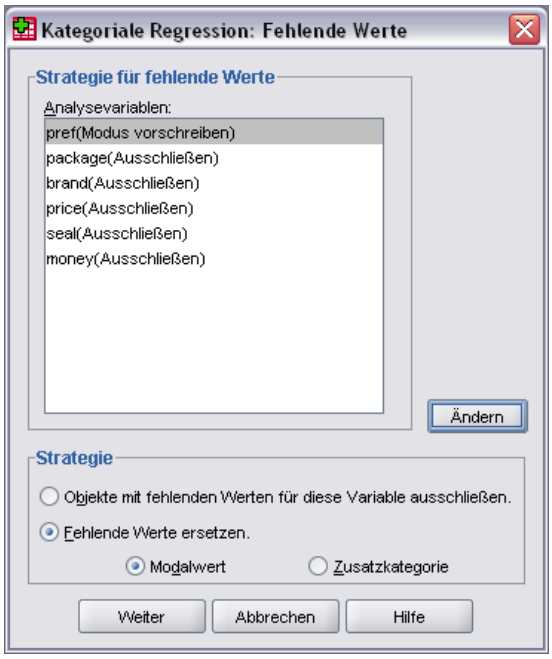
Gruppieren. Folgende Optionen sind bei der Diskretisierung von Variablen durch Gruppierung verfügbar:

- **Anzahl Kategorien.** Geben Sie eine Anzahl von Kategorien an. Legen Sie außerdem fest, ob die Werte der Variablen über die Kategorien ungefähr normal- oder gleichverteilt werden sollen.
- **Gleiche Intervalle.** Die Variablen werden in Kategorien umkodiert, die durch diese gleich großen Intervalle definiert werden. Sie müssen die Größe der Intervalle angeben.

Fehlende Werte in kategorialer Regression

Im Dialogfeld "Kategoriale Hauptkomponenten: Fehlende Werte" können Sie die Strategie für die Verarbeitung von fehlenden Werten in den Analyse- und Zusatzvariablen festlegen.

Abbildung 2-4
Dialogfeld "Fehlende Werte"



Strategie. Legen Sie fest, ob Objekte mit fehlenden Werten ausgeschlossen werden sollen (listenweiser Ausschluss) oder ob fehlende Werte vorgeschrieben (Aktivbehandlung).

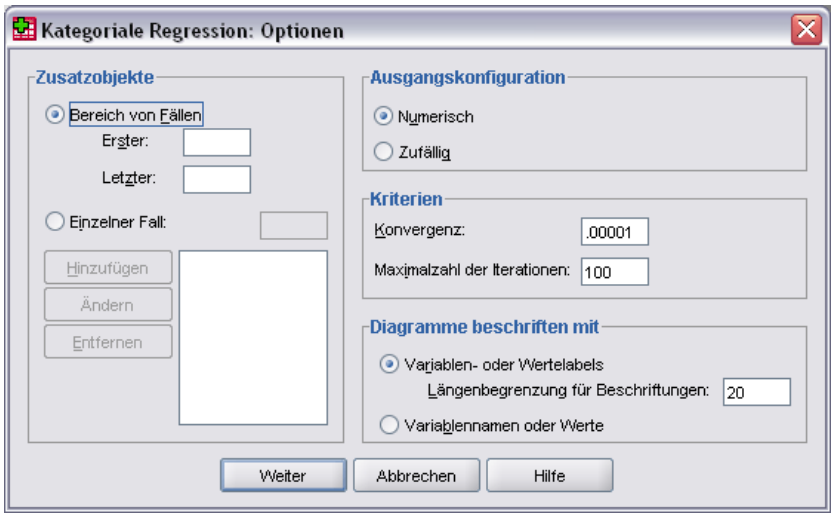
- **Objekte mit fehlenden Werten für diese Variable ausschließen.** Objekte mit fehlenden Werten für die ausgewählte Variable werden aus der Analyse ausgeschlossen. Diese Strategie ist für Zusatzvariablen nicht verfügbar.
- **Fehlende Werte ersetzen.** Fehlende Werte werden bei Objekten mit fehlenden Werten für die ausgewählte Variable gesetzt. Sie können die Methode auswählen, nach der die Werte gesetzt werden: Wählen Sie Modalwert, wenn fehlende Werte durch die häufigste Kategorie ersetzt werden sollen. Falls mehrere Modalwerte vorhanden sind, wird der Modalwert mit der kleinsten Kategorienummer verwendet. Wenn Sie Zusatzkategorie auswählen, werden fehlende Werte durch die gleiche Quantifikation einer zusätzlichen Kategorie ersetzt. Dies

bedeutet aber auch, dass Objekte mit einem fehlenden Wert für diese Variable als zur gleichen Kategorie (der zusätzlichen) zugehörig betrachtet werden.

Kategoriale Regression: Optionen

Im Dialogfeld "Kategoriale Regression: Optionen" können Sie die Ausgangskonfiguration auswählen, Kriterien für die Iterationen und die Konvergenz festlegen, Zusatzobjekte auswählen und die Beschriftung der Diagramme definieren.

Abbildung 2-5
Dialogfeld "Optionen"



Zusatzobjekte. Mit dieser Option können Sie die Objekte angeben, die als Zusatzobjekte behandelt werden sollen. Geben Sie die Nummer eines Zusatzobjekts ein und klicken Sie auf Hinzufügen. Zusatzobjekte können nicht gewichtet werden (Gewichtungen werden ignoriert).

Ausgangskonfiguration. Wenn keine Variablen als nominal behandelt werden, wählen Sie die Konfiguration Numerisch. Wenn mindestens eine Variable als nominal behandelt wird, wählen Sie die Konfiguration Zufällig.

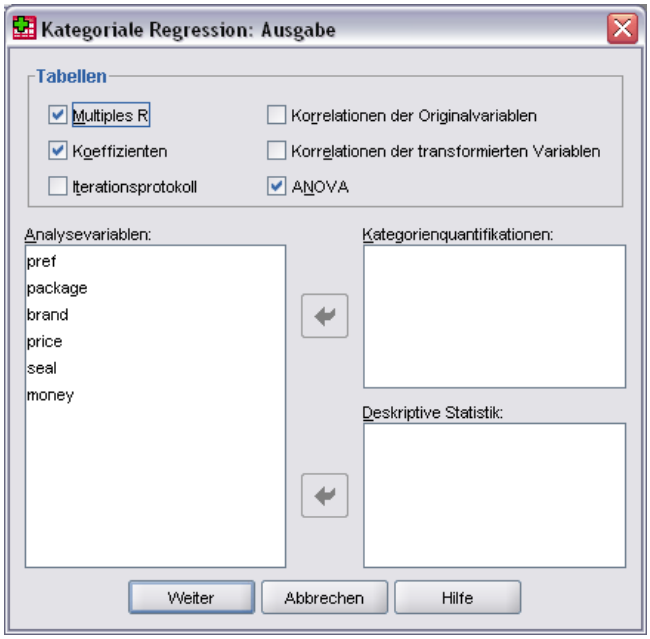
Kriterien. Sie können die Höchstzahl an Iterationen festlegen, die bei den Regressionsberechnungen durchlaufen werden sollen. Außerdem können Sie einen Wert für das Konvergenzkriterium auswählen. Die Regression wird abgebrochen, wenn die Differenz der Gesamtanpassung zwischen den letzten zwei Iterationen kleiner als der Konvergenzwert ist oder die Maximalzahl von Iterationen erreicht wurde.

Diagramme beschriften mit. Hiermit können Sie festlegen, ob in den Diagrammen Variablen- und Wertlabels oder Variablennamen und Werte verwendet werden sollen. Sie können auch eine Höchstlänge für die Beschriftungen festlegen.

Kategoriale Regression: Ausgabe

Im Dialogfeld "Ausgabe" können Sie die Statistiken auswählen, die in der Ausgabe angezeigt werden sollen.

Abbildung 2-6
Dialogfeld "Ausgabe"



Tabellen. Es werden die folgenden Tabellen erstellt:

- **Multiples R.** Umfaßt R^2, korrigiertes R^2 und korrigiertes R^2 unter Berücksichtigung der optimalen Skalierung.
- **Koeffizienten.** Diese Option gibt drei Tabellen aus: eine Koeffiziententabelle mit Beta-Koeffizienten, dem Standardfehler der Beta-Koeffizienten, *t*-Werten und der Signifikanz; eine Tabelle der optimalen Skalierung der Koeffizienten mit dem Standardfehler der Beta-Koeffizienten unter Berücksichtigung der optimalen Skalierungsfreiheitsgrade; und eine Tabelle mit Korrelationen nullter Ordnung, Teilkorrellationen und partiellen Korrelationen, Pratts Maß für die relative Bedeutung der transformierten Einflußvariablen und der Toleranz vor und nach der Transformation.
- **Iterationsprotokoll.** Für jede Iteration, einschließlich der Anfangswerte des Algorithmus, werden das multiple *R* und der Regressionsfehler angezeigt. Die Erhöhung des multiplen *R* wird beginnend mit der ersten Iteration aufgeführt.
- **Korrelationen der Originalvariablen.** Eine Matrix mit den Korrelationen zwischen den nicht transformierten Variablen wird angezeigt.
- **Korrelationen der transformierten Variablen.** Eine Matrix mit den Korrelationen zwischen den transformierten Variablen wird angezeigt.
- **ANOVA.** Diese Option enthält Quadratsummen für die Regression und die Residuen, Mittel der Quadrate und *F*. Zwei ANOVA-Tabellen werden angezeigt: eine Tabelle mit Freiheitsgraden für die Regression, die der Anzahl der Einflußvariablen entspricht, und eine Tabelle mit Freiheitsgraden für die Regression unter Berücksichtigung der optimalen Skalierung.

Kategorienquantifikationen. Tabellen mit den transformierten Werten der gewählten Variablen werden angezeigt.

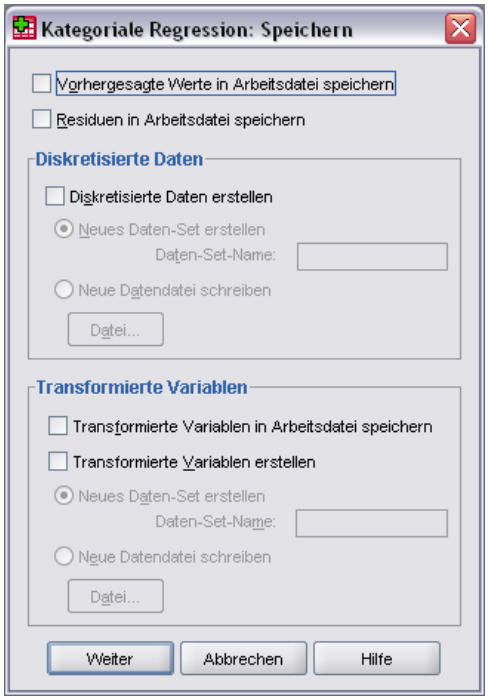
Deskriptive Statistiken. Tabellen mit den Häufigkeiten, fehlenden Werten und Modalwerten der gewählten Variablen werden angezeigt.

Kategoriale Regression: Speichern

Im Dialogfeld "Speichern" können Sie festlegen, dass die vorhergesagten Werte, Residuen und transformierten Werte in der Arbeitsdatei gespeichert werden. Außerdem können die diskretisierten Daten und die transformierten Werte in einer SPSS-Datendatei oder einem Daten-Set gespeichert werden.

- Daten-Sets sind in der aktuellen Sitzung verfügbar. In nachfolgenden Sitzungen stehen Daten-Sets nur zur Verfügung, wenn Sie sie ausdrücklich als Datendateien speichern. Die Namen von Daten-Sets müssen den Regeln zum Benennen von Variablen entsprechen.
- Die Namen für die Dateien und Daten-Sets müssen sich bei allen Arten von gespeicherten Daten unterscheiden.

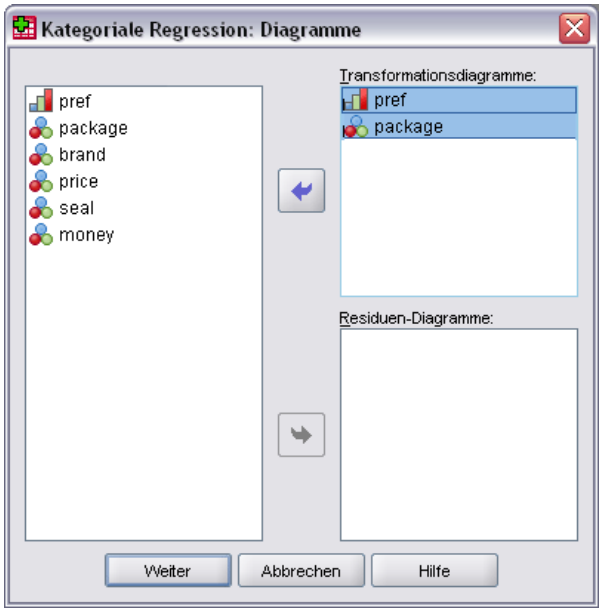
Abbildung 2-7
Dialogfeld "Speichern"



Kategoriale Regression: Transformationsdiagramme

Im Dialogfeld "Diagramme" können Sie die Variablen angeben, aus denen Transformationsdiagramme und Residuen-Diagramme erzeugt werden.

Abbildung 2-8
Dialogfeld "Diagramme"



Transformationsdiagramme. Für jede dieser Variablen werden die Kategorienquantifikationen anhand der ursprünglichen Kategoriewerte dargestellt. Leere Kategorien erscheinen auf der horizontalen Achse, haben jedoch keine Auswirkungen auf die Berechnungen. Diese Kategorien werden durch Unterbrechungen in der Verbindungslinie zwischen den Quantifikationen gekennzeichnet.

Residuen-Diagramme. Für jede dieser Variablen werden Residuen (berechnet für die abhängige Variable, die aus allen Einflußvariablen mit Ausnahme der betreffenden Einflußvariable prognostiziert wurde) anhand der Kategorienummern und der optimalen Kategorienquantifikationen multipliziert mit dem Beta über Kategorienummern grafisch dargestellt.

Zusätzliche Funktionen beim Befehl CATREG

Sie können die kategoriale Regression Ihren Anforderungen entsprechend anpassen, wenn Sie ihre Auswahl in ein Syntax-Fenster einfügen und die resultierende Befehlssyntax für den Befehl CATREG bearbeiten. Mit der Befehlssyntax-Sprache verfügen Sie außerdem über folgende Möglichkeiten:

- Mit dem Unterbefehl SAVE können Sie Stammnamen für die transformierten Variablen festlegen, wenn Sie diese in der Arbeitsdatei speichern.

Vollständige Informationen zur Syntax finden Sie in der *Command Syntax Reference*.

Kapitel
3

Kategoriale Hauptkomponentenanalyse (CATPCA)

Bei diesem Verfahren werden kategoriale Variablen quantifiziert und gleichzeitig die Dimensionen der Daten reduziert. Die kategoriale Hauptkomponentenanalyse ist auch unter dem Akronym CATPCA bekannt (*Categorical Principal Components Analysis*).

Das Ziel bei der Hauptkomponentenanalyse besteht darin, ein ursprüngliches Set von Variablen in ein kleineres Set von unkorrelierten Komponenten umzuwandeln, das soviel wie möglich von den Informationen aus den ursprünglichen Variablen enthält. Diese Methode ist besonders dann nützlich, wenn eine hohe Anzahl von Variablen eine effektive Interpretation der Beziehungen zwischen den Objekten (Subjekten und Einheiten) verhindert. Nach einer Dimensionsreduktion werden nur einige Komponenten anstelle einer großen Anzahl von Variablen interpretiert.

Beim Standardverfahren der Hauptkomponentenanalyse wird von linearen Beziehungen zwischen numerischen Variablen ausgegangen. Bei der Methode der optimalen Skalierung können die Variablen jedoch auf verschiedenen Niveaus skaliert werden. Kategoriale Variablen werden optimal in der angegebenen Dimensionalität quantifiziert. Aufgrund dessen können nichtlineare Beziehungen zwischen Variablen modelliert werden.

Beispiel. Eine kategoriale Hauptkomponentenanalyse ließe sich verwenden, um den Zusammenhang zwischen Berufsgruppe, beruflicher Stellung, Region, Anzahl von Dienstreisen (hoch, mittel und gering) und der Zufriedenheit im Beruf grafisch darzustellen. Das Ergebnis könnte sein, dass ein Großteil der Varianz mit zwei Dimensionen erklärt werden kann. In der ersten Dimension ließen sich Berufsgruppe von Region trennen, und in der zweiten Dimension könnten berufliche Stellung und Häufigkeit von Dienstreisen getrennt werden. Außerdem könnte sich ein Zusammenhang zwischen hoher Zufriedenheit im Beruf und einer mittleren Häufigkeit von Dienstreisen ergeben.

Statistiken und Diagramme. Häufigkeiten, fehlende Werte, Messniveau der optimalen Skalierung, Modalwert, berücksichtige Varianz nach Koordinaten der Zentroiden, Vektorkoordinaten, Gesamtsumme pro Variable und Dimension, Komponentenladungen für vektorquantifizierte Variablen, Quantifikationen und Koordinaten der Kategorien, Iterationsprotokoll, Korrelationen der transformierten Variablen und Eigenwerte der Korrelationsmatrix, Korrelationen der ursprünglichen Variablen und Eigenwerte der Korrelationsmatrix, Objektwerte, Diagramme der Kategorien und verknüpften Kategorien, Transformationsdiagramme, Residuen-Diagramme, Diagramme der projizierten Zentroide, Objektdiagramme, Biplots, Triplots und Diagramme der Komponentenladungen.

Daten. Werte von String-Variablen werden immer in aufsteigender alphanumerischer Reihenfolge in positive Ganzzahlen umgewandelt. Benutzerdefinierte und systemdefinierte fehlende Werte sowie Werte kleiner als 1 werden als fehlend betrachtet. Sie können die Daten umkodieren oder eine Konstante zu Variablen mit Werten kleiner als 1 addieren, um die Werte als nichtfehlend zu deklarieren.

Annahmen. Die Daten müssen mindestens drei gültige Fälle enthalten. Die Analyse basiert auf Daten, die aus positiven ganzen Zahlen bestehen. Mit der Option zur Diskretisierung werden Variablen mit Dezimalwerten automatisch kategorisiert, indem die Werte in Kategorien gruppiert werden, die nahezu normalverteilt sind, und die Werte von String-Variablen automatisch in positive Ganzzahlen umgewandelt. Sie können andere Diskretisierungsmethoden bestimmen.

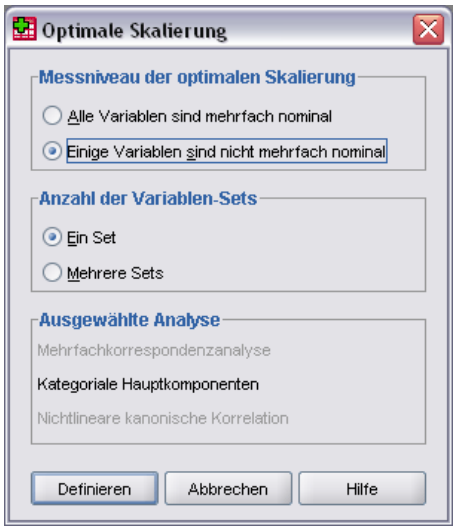
Verwandte Prozeduren. Eine Skalierung aller Variablen auf numerischem Niveau entspricht der normalen Hauptkomponentenanalyse. Wenn Sie die transformierten Variablen in einer normalen linearen Hauptkomponentenanalyse verwenden, stehen Ihnen andere Diagrammfunktionen zur Verfügung. Wenn alle Variablen ein mehrfach nominales Skalierungsniveau besitzen, ist die kategoriale Hauptkomponentenanalyse identisch mit der Mehrfachkorrespondenzanalyse. Wenn Variablen-Sets untersucht werden sollen, verwenden Sie die kategoriale (nichtlineare) kanonische Korrelationsanalyse.

So lassen Sie eine kategoriale Hauptkomponentenanalyse berechnen:

- Wählen Sie die folgenden Befehle aus den Menüs aus:

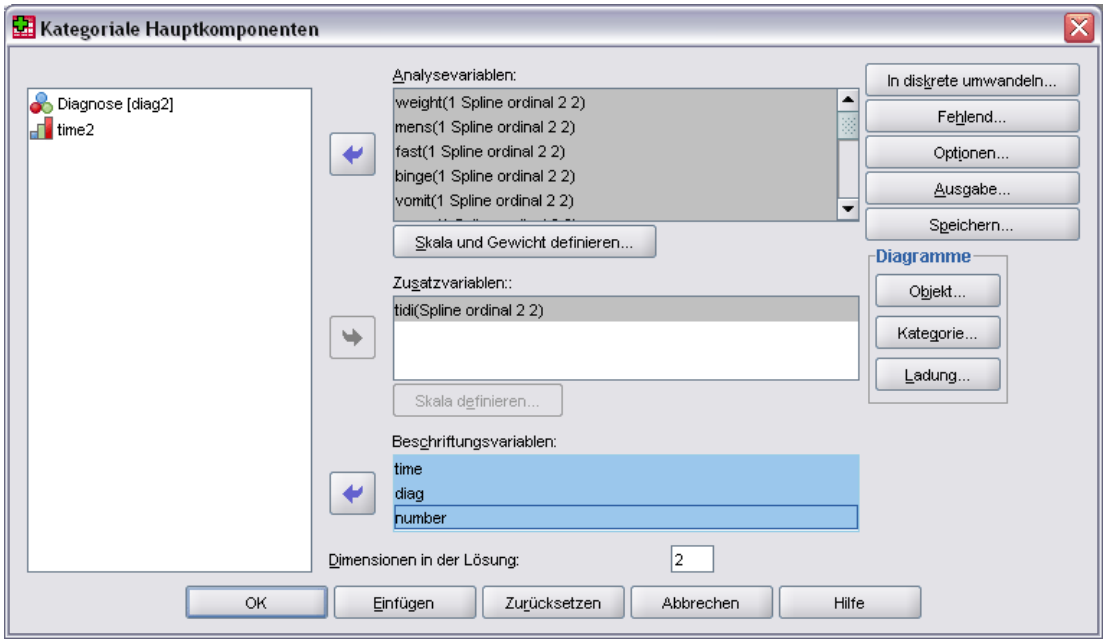
  Analysieren
  Dimensionsreduktion
  Optimale Skalierung...

Abbildung 3-1
Dialogfeld "Optimale Skalierung"



- Wählen Sie Einige Variablen sind nicht mehrfach nominal aus.
- Wählen Sie die Option Ein Set.
- Klicken Sie auf Definieren.

Abbildung 3-2
Dialogfeld "Kategoriale Hauptkomponenten"



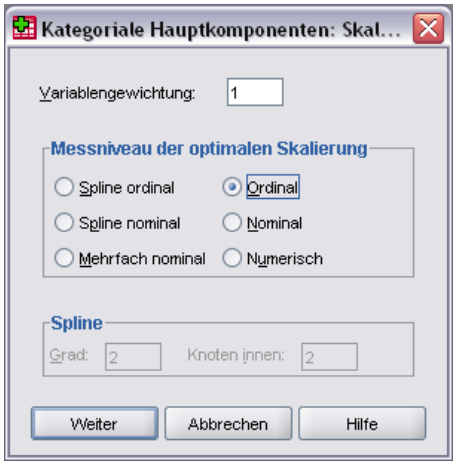
- Wählen Sie mindestens zwei Analysevariablen aus und geben Sie die Anzahl der Dimensionen in der Lösung an.
- Klicken Sie auf OK.

  Sie können außerdem Beschriftungsvariablen für die Diagramme und Zusatzvariablen bestimmen, die in die gefundene Lösung eingepaßt werden.

Skala und Gewicht definieren in CATPCA

Sie können das Messniveau der optimalen Skalierung für die Analysevariablen und die Zusatzvariablen festlegen. In der Standardeinstellung werden sie als monotone Splines zweiten Grades (ordinal) mit zwei inneren Knoten skaliert. Außerdem können Sie die Gewichtung der Analysevariablen bestimmen.

Abbildung 3-3
Dialogfeld "Skala und Gewichtung definieren"



Variablengewichtung. Sie können für jede Variable eine Gewichtung festlegen. Bei dem angegebenen Wert muss es sich um eine positive ganze Zahl handeln. Der Standardwert ist 1.

Messniveau der optimalen Skalierung. Sie können auch das für die Quantifikation der einzelnen Variablen verwendete Skalierungsniveau auswählen.

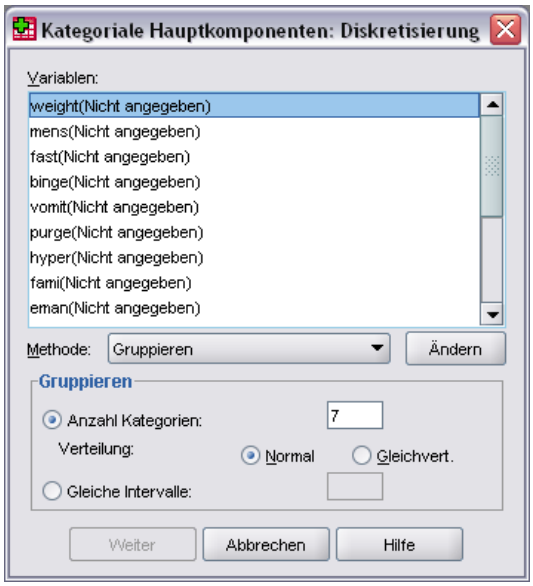
- **Spline ordinal.** Die Reihenfolge der Kategorien bei der beobachteten Variable wird bei der optimal skalierten Variable beibehalten. Die Kategorienpunkte werden auf einer geraden Linie (einem Vektor) durch den Koordinatenursprung plaziert. Die resultierende Transformation ist ein glatter, stückweiser, monotoner Polynom mit dem gewählten Grad. Die Stücke werden durch die vom Benutzer festgelegte Anzahl und die durch die Prozedur bestimmte Lage der inneren Knoten definiert.
- **Spline nominal.** Die einzige Information aus der beobachteten Variable, die in der optimal skalierten Variable beibehalten wird, ist die Gruppierung der Objekte in Kategorien. Die Reihenfolge der Kategorien bei der beobachteten Variablenwird nicht beibehalten. Die Kategorienpunkte werden auf einer geraden Linie (einem Vektor) durch den Koordinatenursprung plaziert. Die resultierende Transformation ist ein glatter, stückweiser, möglicherweise nichtmonotoner Polynom mit dem gewählten Grad. Die Stücke werden durch die vom Benutzer festgelegte Anzahl und die durch die Prozedur bestimmte Lage der inneren Knoten definiert.
- **Mehrfach nominal.** Die einzige Information aus der beobachteten Variable, die in der optimal skalierten Variable beibehalten wird, ist die Gruppierung der Objekte in Kategorien. Die Reihenfolge der Kategorien bei der beobachteten Variablenwird nicht beibehalten. Die Kategoriepunkte werden in den Zentroiden der Objekte in den einzelnen Kategorien plaziert. Die Bezeichnung *Mehrfach* bedeutet, dass für jede Dimension andere Gruppen von Quantifikationen berechnet werden.
- **Ordinal.** Die Reihenfolge der Kategorien bei der beobachteten Variable wird bei der optimal skalierten Variable beibehalten. Die Kategorienpunkte werden auf einer geraden Linie (einem Vektor) durch den Koordinatenursprung plaziert. Die resultierende Transformation ist besser angepaßt, aber weniger glatt als bei "Spline ordinal".

- **Nominal.** Die einzige Information aus der beobachteten Variable, die in der optimal skalierten Variable beibehalten wird, ist die Gruppierung der Objekte in Kategorien. Die Reihenfolge der Kategorien bei der beobachteten Variablenwird nicht beibehalten. Die Kategorienpunkte werden auf einer geraden Linie (einem Vektor) durch den Koordinatenursprung plaziert. Die resultierende Transformation ist besser angepaßt, aber weniger glatt als bei "Spline nominal".
- **Numerisch.** Die Kategorien werden als geordnet mit gleichen Abständen behandelt (Intervallniveau). Die Reihenfolge der Kategorien und die gleichen Abstände zwischen den Kategorienummern in der beobachteten Variablenbleiben in der optimal skalierten Variablen erhalten. Die Kategorienpunkte werden auf einer geraden Linie (einem Vektor) durch den Koordinatenursprung plaziert. Wenn alle Variablen auf numerischem Niveau skaliert sind, entspricht die Analyse der normalen Hauptkomponentenanalyse.

Kategoriale Hauptkomponenten: Diskretisierung

Im Dialogfeld "Kategoriale Hauptkomponenten: Diskretisierung" können Sie eine Methode zum Umkodieren der Variablen festlegen. Sofern nicht anders angegeben, werden Variablen mit Dezimalwerten in sieben Kategorien gruppiert, die nahezu normalverteilt sind. (Wenn die Variable weniger als sieben unterschiedliche Werte aufweist, entspricht die Anzahl der Kategorien der Anzahl der unterschiedlichen Werte.) Werte von String-Variablen werden immer in positive ganze Zahlen umgewandelt, indem in aufsteigender alphanumerischer Reihenfolge Kategorienummern zugewiesen werden. Die Diskretisierung von String-Variablen erfolgt anhand dieser Ganzzahlen. Andere Variablen werden in der Standardeinstellung nicht modifiziert. Die diskretisierten Variablen werden dann in der Analyse verwendet.

Abbildung 3-4
Dialogfeld "Diskretisierung"



Methode. Verfügbar sind die Optionen "Gruppieren", "Einstufen" oder "Multiplizieren".

- **Gruppieren.** Hierbei werden die Daten in eine bestimmte Anzahl von Kategorien oder nach Intervall umkodiert.

- **Einstufen.** Die Variable wird durch Bilden einer Rangfolge diskretisiert.
- **Multiplizieren.** Die aktuellen Werte der Variablen werden standardisiert, mit 10 multipliziert und gerundet. Anschließend wird eine Konstante zu den Werten addiert, sodass der niedrigste diskretisierte Wert 1 beträgt.

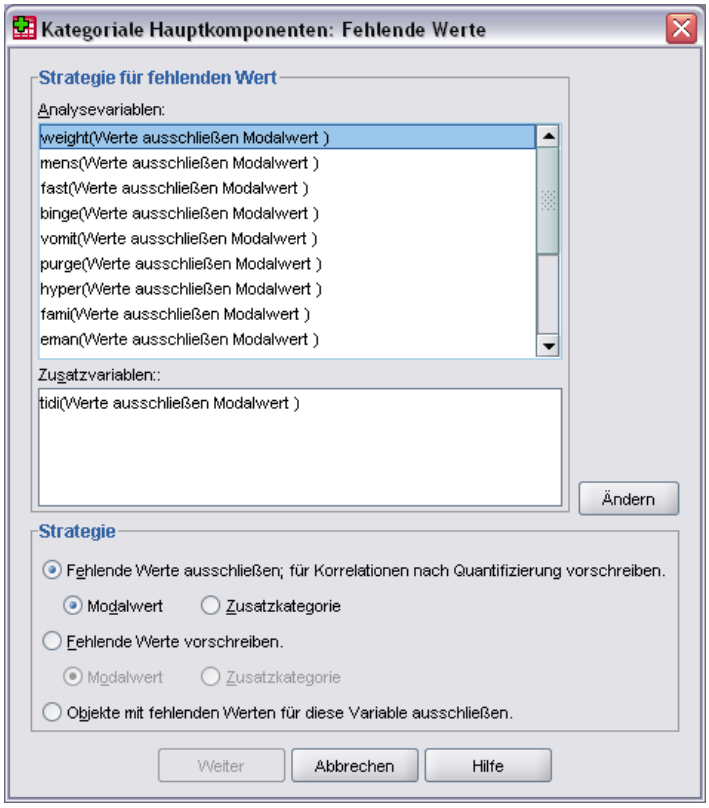
Gruppieren. Folgende Optionen sind bei der Diskretisierung von Variablen durch Gruppierung verfügbar:

- **Anzahl Kategorien.** Geben Sie eine Anzahl von Kategorien an. Legen Sie außerdem fest, ob die Werte der Variablen über die Kategorien ungefähr normal- oder gleichverteilt werden sollen.
- **Gleiche Intervalle.** Die Variablen werden in Kategorien umkodiert, die durch diese gleich großen Intervalle definiert werden. Sie müssen die Größe der Intervalle angeben.

Kategoriale Hauptkomponenten: Fehlende Werte

Im Dialogfeld "Kategoriale Hauptkomponenten: Fehlende Werte" können Sie die Strategie für die Verarbeitung von fehlenden Werten in den Analyse- und Zusatzvariablen festlegen.

Abbildung 3-5
Dialogfeld "Fehlende Werte"



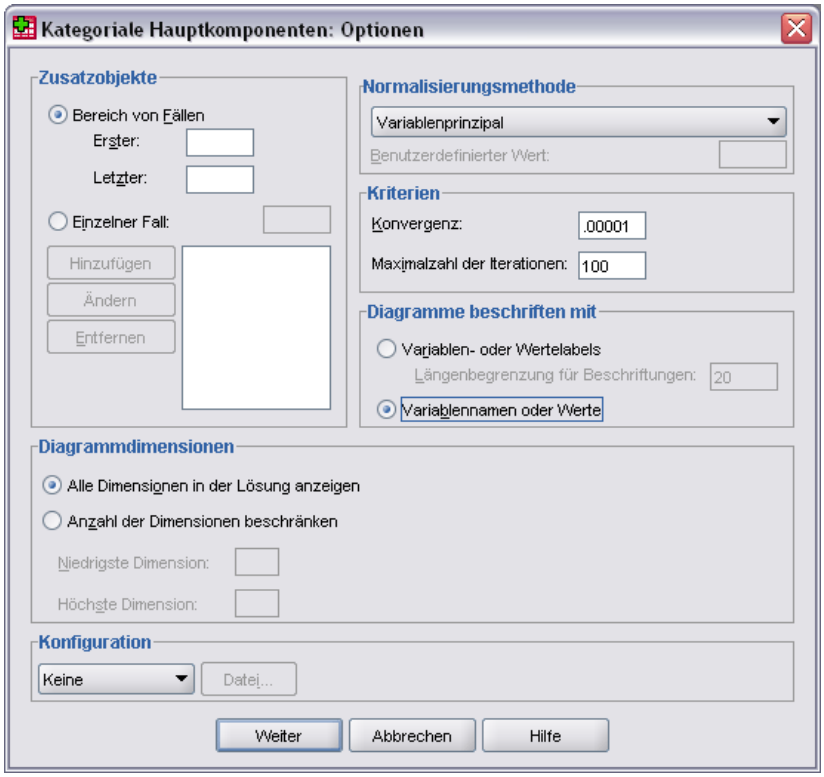
Strategie. Fehlende Werte können ausgeschlossen (Passivbehandlung) oder vorgeschrieben (Aktivbehandlung) werden oder Objekte mit fehlenden Werten können ausgeschlossen werden (listenweiser Fallausschluss).

- **Fehlende Werte ausschließen; für Korrelationen nach Quantifizierung vorschreiben.** Objekte mit fehlenden Werten bei der ausgewählten Variablen tragen nicht zur Analyse dieser Variablen bei. Wenn alle Variablen passiv behandelt werden, werden Objekte mit fehlenden Werten für alle Variablen als Zusatzobjekte behandelt. Wenn im Dialogfeld "Kategoriale Hauptkomponenten: Ausgabe" Korrelationen ausgewählt sind, werden fehlende Werte nach der Analyse anhand der häufigsten Variablenkategorie (Modalwert) für die Korrelationen der ursprünglichen Variablengesetzt. Sie können die Methode auswählen, nach der Werte für die Korrelationen der optimal skalierten Variablen gesetzt werden. Wenn Sie Modalwert auswählen, werden fehlende Werte durch den Modalwert der optimal skalierten Variablen ersetzt. Wenn Sie Zusatzkategorie auswählen, werden fehlende Werte durch die Quantifikation einer zusätzlichen Kategorie ersetzt. Dies bedeutet aber auch, dass Objekte mit einem fehlenden Wert für diese Variable als zur gleichen Kategorie (der zusätzlichen) zugehörig betrachtet werden.
- **Fehlende Werte ersetzen.** Fehlende Werte werden bei Objekten mit fehlenden Werten für die ausgewählte Variable gesetzt. Sie können die Methode auswählen, nach der die Werte gesetzt werden: Wählen Sie Modalwert, wenn fehlende Werte durch die häufigste Kategorie ersetzt werden sollen. Falls mehrere Modalwerte vorhanden sind, wird der Modalwert mit der kleinsten Kategorienummer verwendet. Wenn Sie Zusatzkategorie auswählen, werden fehlende Werte durch die gleiche Quantifikation einer zusätzlichen Kategorie ersetzt. Dies bedeutet aber auch, dass Objekte mit einem fehlenden Wert für diese Variable als zur gleichen Kategorie (der zusätzlichen) zugehörig betrachtet werden.
- **Objekte mit fehlenden Werten für diese Variable ausschließen.** Objekte mit fehlenden Werten für die ausgewählte Variable werden aus der Analyse ausgeschlossen. Diese Strategie ist für Zusatzvariablen nicht verfügbar.

Kategoriale Hauptkomponenten: Optionen

Im Dialogfeld "Kategoriale Hauptkomponenten: Optionen" können Sie die anfängliche Konfiguration auswählen, Kriterien für die Iterationen und die Konvergenz festlegen, eine Normalisierungsmethode auswählen, die Methode zum Beschriften der Diagramme auswählen und Zusatzobjekte angeben.

Abbildung 3-6
Dialogfeld "Optionen"



Zusatzobjekte. Geben Sie die Fallnummer des Objekts bzw. die erste und letzte Fallnummer einer Reihe von Objekten an, die als Zusatzobjekte deklariert werden sollen, und klicken Sie dann auf Hinzufügen. Wiederholen Sie diesen Vorgang, bis Sie alle Zusatzobjekte festgelegt haben. Wenn ein Objekt als Zusatzobjekt deklariert ist, werden Fallgewichtungen für das Objekt ignoriert.

Normalisierungsmethode. Sie können eine von fünf Optionen für die Normalisierung der Objektwerte und Variablen auswählen. In einer Analyse kann nur eine Normalisierungsmethode gleichzeitig verwendet werden.

- **Variablenprinzipal.** Bei dieser Option wird die Assoziation zwischen Variablen optimiert. Die Komponentenladungen (Korrelationen mit Hauptkomponenten wie Dimensionen und Objektwerten) stellen die Koordinaten der Variablen im Objektraum dar. Dies ist nützlich, wenn Sie hauptsächlich an der Korrelation zwischen den Variablen interessiert sind.
- **Objektprinzipal.** Bei dieser Option werden die Distanzen zwischen den Objekten optimiert. Dies ist nützlich, wenn Sie hauptsächlich an den Unterschieden oder Übereinstimmungen zwischen den Objekten interessiert sind.
- **Symmetrisch.** Verwenden Sie diese Normalisierungsmethode, wenn Sie hauptsächlich an den Beziehungen zwischen den Objekten und den Variablen interessiert sind.

- **Unabhängig.** Verwenden Sie diese Normalisierungsmethode, wenn Sie die Distanzen zwischen den Objekten und die Korrelationen zwischen den Variablen getrennt untersuchen möchten.
- **Benutzerdefiniert.** Sie können einen beliebigen reellen Wert aus dem abgeschlossenen Intervall [–1, 1] eingeben. Der Wert 1 entspricht der Methode "Objektprinzipal", der Wert 0 der Methode "Symmetrisch" und der Wert –1 der Methode "Variablenprinzipal". Bei Werten größer als –1 und kleiner als 1 wird der Eigenwert sowohl über Objekte als auch Variablen gestreut. Mit dieser Methode können Sie Biplots und Triplots genau an Ihre Bedürfnisse anpassen.

Kriterien. Sie können die maximale Anzahl von Iterationen festlegen, die in der Prozedur bei den Berechnungen durchlaufen werden. Außerdem können Sie einen Wert für das Konvergenzkriterium auswählen. Die Prozedur wird abgebrochen, wenn die Differenz der Gesamtanpassung zwischen den letzten zwei Iterationen kleiner als der Konvergenzwert ist oder die Maximalzahl von Iterationen erreicht wurde.

Diagramme beschriften mit. Hiermit können Sie festlegen, ob in den Diagrammen Variablen- und Wertlabels oder Variablennamen und Werte verwendet werden sollen. Sie können auch eine Höchstlänge für die Beschriftungen festlegen.

Diagrammdimensionen. Hiermit legen Sie fest, wie die Statistiken in der Ausgabe angezeigt werden.

- **Alle Dimensionen in der Lösung anzeigen.** Alle Dimensionen in der Lösung werden in einer Streudiagramm-Matrix angezeigt.
- **Anzahl der Dimensionen beschränken.** Die angezeigten Dimensionen werden auf geplottete Paare beschränkt. Wenn Sie die Dimensionen beschränken, müssen Sie die niedrigste und die höchste Dimension auswählen, die jeweils geplottet werden soll. Die niedrigste Dimension muss im Bereich von 1 bis zur Anzahl der Dimensionen in der Lösung minus 1 liegen und wird gegen höhere Dimensionen geplottet. Der Wert für die höchste Dimension muss im Bereich von 2 bis zur Anzahl der Dimensionen in der Lösung liegen und gibt die höchste Dimension an, die beim Plotten der Dimensionspaare verwendet wird. Diese Spezifikation gilt für alle angeforderten multidimensionalen Plots.

Konfiguration. Sie können die Koordinaten einer Konfiguration aus einer Datei einlesen. Die erste Variable in der Datei muss die Koordinaten der ersten Dimension, die zweite Variable die Koordinaten der zweiten Dimension usw. enthalten.

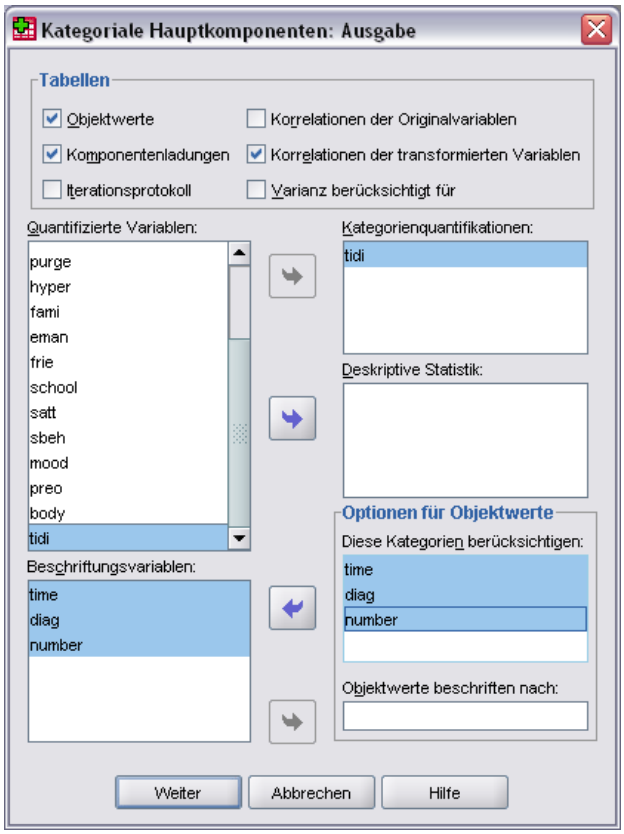
- **Initial.** Die in der Datei angegebene Konfiguration wird als Ausgangspunkt für die Analyse verwendet.
- **Fest.** Die in der Datei angegebene Konfiguration wird zum Anpassen der Variablen verwendet. Die Variablen, die angepaßt werden, müssen als Analysevariablen ausgewählt werden. Da die Konfiguration aber fest ist, werden Sie als Zusatzvariablen verarbeitet (und müssen deshalb nicht als Zusatzvariablen ausgewählt werden).

Kategoriale Hauptkomponenten: Ausgabe

Im Dialogfeld "Kategoriale Hauptkomponenten: Ausgabe" können Sie festlegen, dass Tabellen für Objektwerte, Komponentenladungen, das Iterationsprotokoll, Korrelationen der ursprünglichen und transformierten Variablen, berücksichtigte Varianz pro Variable und

Dimension, Kategorienquantifikationen für ausgewählte Variablen und deskriptive Statistiken für ausgewählte Variablen erstellt werden sollen.

Abbildung 3-7
Dialogfeld "Ausgabe"



Objektwerte. Hiermit werden die Objektwerte angezeigt. Die folgenden Optionen sind verfügbar:

- **Diese Kategorien berücksichtigen.** Hiermit werden die Kategorienummern der ausgewählten Analysevariablen angezeigt.
- **Objektwerte beschriften nach.** Sie können eine der in der Liste Beschriftungsvariablen enthaltenen Variablen auswählen, um die Objekte zu beschriften.

Komponentenladungen. Hiermit werden Komponentenladungen für alle Variablen angezeigt, bei denen keine mehrfach nominalen Skalierungsniveaus vorliegen.

Iterationsprotokoll. Für jede Iteration werden berücksichtigte Varianz, Verlust sowie Steigerung der berücksichtigten Varianz ausgegeben.

Korrelationen der Originalvariablen. Hiermit werden die Korrelationsmatrix der ursprünglichen Variablen und die Eigenwerte dieser Matrix ausgegeben.

Korrelationen der transformierten Variablen. Hiermit werden die Korrelationsmatrix der transformierten (optimal skalierten) Variablen und die Eigenwerte dieser Matrix ausgegeben.

Varianz berücksichtigt für. Hiermit wird die berücksichtigte Varianz nach Zentroidkoordinaten und nach Vektorkoordinaten sowie die Gesamtsumme (kombinierte Zentroid- und Vektorkoordinaten) pro Variable und Dimension ausgegeben.

Kategorienquantifikationen. Dies sind die Kategoriequantifikationen und -koordinaten pro Dimension jeder ausgewählten Variable.

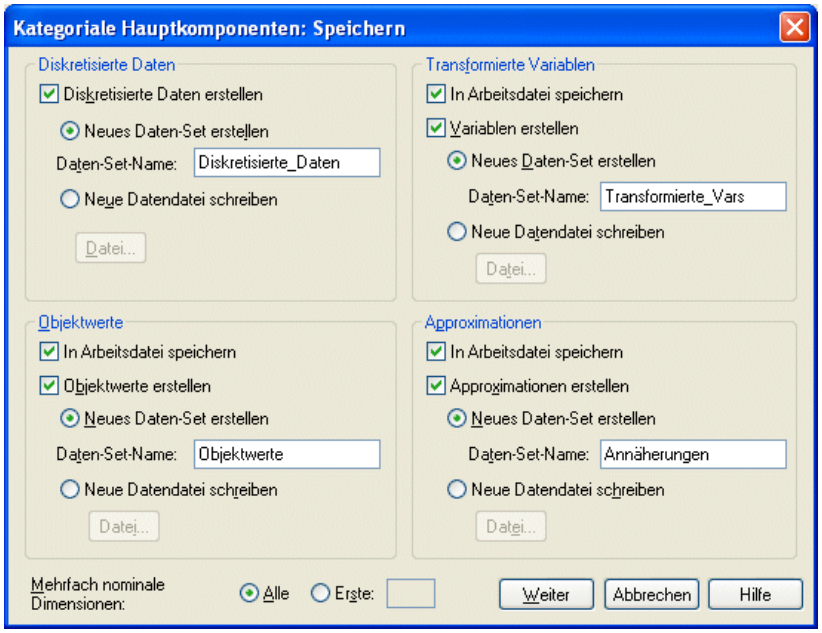
Deskriptive Statistik. Hiermit werden Häufigkeiten, Anzahl der fehlenden Werte und Modalwert für die ausgewählten Variablen ausgegeben.

Kategoriale Hauptkomponenten: Speichern

Im Dialogfeld “Speichern” können Sie festlegen, dass diskretisierte Daten, Objektwerte, transformierte Werte und Approximationen in einer SPSS-Datendatei oder in einem Daten-Set gespeichert werden. Außerdem können die transformierten Werte, die Objektwerte und die Approximationen in der Arbeitsdatei gespeichert werden.

- Daten-Sets sind in der aktuellen Sitzung verfügbar. In nachfolgenden Sitzungen stehen Daten-Sets nur zur Verfügung, wenn Sie sie ausdrücklich als Datendateien speichern. Die Namen von Daten-Sets müssen den Regeln zum Benennen von Variablen entsprechen.
- Die Namen für die Dateien und Daten-Sets müssen sich bei allen Arten von gespeicherten Daten unterscheiden.
- Wenn Sie die Objektwerte oder die transformierten Werte in der Arbeitsdatei speichern, können Sie die Anzahl der mehrfach nominalen Dimensionen angeben.

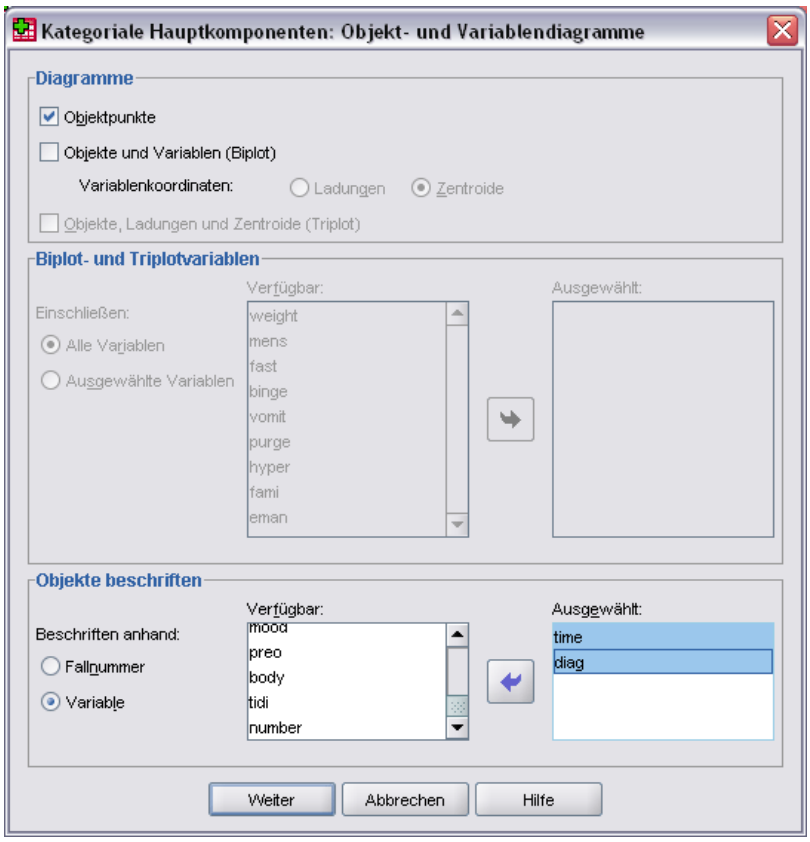
Abbildung 3-8
Dialogfeld “Speichern”



Kategoriale Hauptkomponenten: Objekt- und Variablendiagramme

Im Dialogfeld "Kategoriale Hauptkomponenten: Objekt- und Variablendiagramme" können Sie festlegen, welche Diagramme und für welche Variablen die Diagramme erstellt werden sollen.

Abbildung 3-9
Dialogfeld "Objekt- und Variablendiagramme"



Objektpunkte. Es wird ein Diagramm der Objektpunkte angezeigt.

Objekte und Variablen (Biplot). Es werden Objektpunkte und Variablenkoordinaten ausgegeben. Hierbei bestehen die folgenden Möglichkeiten: Komponentenladungen oder Zentroide der Variablen.

Objekte, Ladungen und Zentroide (Triplot). Es werden die Objektpunkte, die Zentroide von Variablen mit mehrfach nominalem Skalierungsniveau und die Komonentenladungen von anderen Variablen ausgegeben.

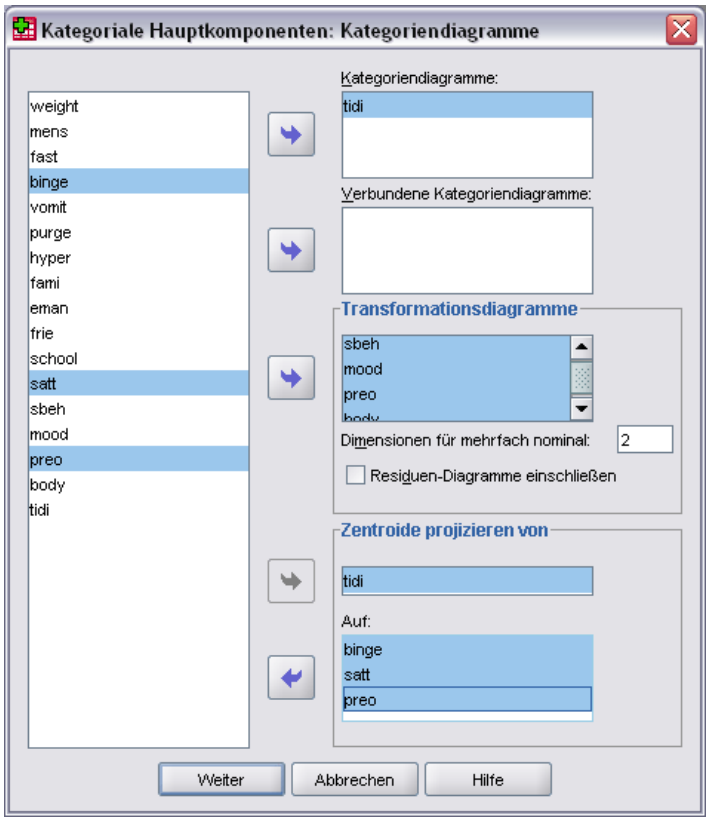
Biplot- und Triplotvariablen. Die Biplots und Triplots können auf allen Variablen oder einer Teilmenge der Variablen basieren.

Objekte beschriften. Objekte können mit den Kategorien ausgewählter Variablen (im Dialogfeld "Kategoriale Hauptkomponenten: Optionen" können Sie Kategorienummern oder Wertelabels wählen) oder den Fallnummern beschriftet werden. Wenn Variable ausgewählt ist, wird pro Variable ein Diagramm erstellt.

Kategoriale Hauptkomponenten: Kategoriendiagramme

Im Dialogfeld “Kategoriale Hauptkomponenten: Kategoriendiagramme” können Sie festlegen, welche Diagramme und für welche Variablen die Diagramme erstellt werden sollen.

Abbildung 3-10
Dialogfeld “Kategoriendiagramme”



Kategoriendiagramme. Für jede ausgewählte Variable wird ein Diagramm mit Zentroid und Vektorkoordinaten erstellt. Bei Variablen mit mehrfach nominalem Skalierungsniveau befinden sich die Kategorien in den Zentroiden der Objekte in der jeweiligen Kategorie. Bei allen anderen Skalierungsniveaus befinden sich die Kategorien auf einem Vektor, der durch den Koordinatenursprung verläuft.

Verbundene Kategoriendiagramme. Hierbei handelt es sich um ein einzelnes Diagramm mit Zentroid und Vektorkoordinaten jeder ausgewählten Variablen.

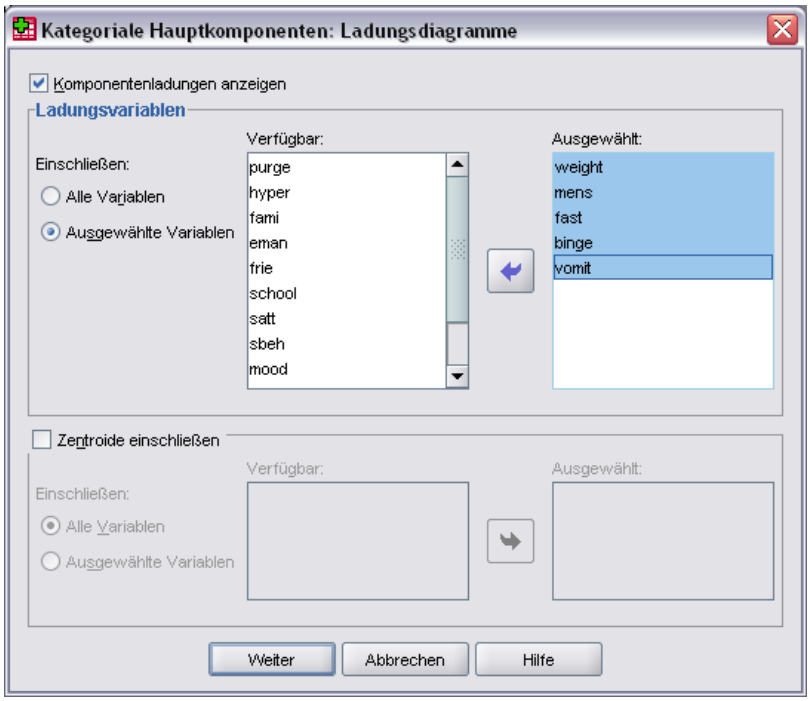
Transformationsdiagramme. Hierbei wird ein Diagramm der optimalen Kategorienquantifikationen gegen die Kategorienummern erstellt. Bei Variablen mit mehrfach nominalem Skalierungsniveau können Sie die gewünschte Anzahl von Dimensionen angeben. Pro Dimension wird ein Diagramm erstellt. Sie können außerdem festlegen, dass für jede ausgewählte Variable ein Diagramm der Residuen ausgegeben wird.

Zentroide projizieren von. Sie können eine Variable auswählen und ihre Zentroide auf ausgewählte Variablen projizieren. Es kann jedoch nicht auf Variablen mit mehrfach nominalem Skalierungsniveau projiziert werden. Wenn dieses Diagramm angefordert wird, wird außerdem eine Tabelle mit den Koordinaten der projizierten Zentroiden ausgegeben.

Kategoriale Hauptkomponentenanalyse: Ladungsdiagramme

Im Dialogfeld "Ladungsdiagramme" können Sie die Variablen angeben, die in das Diagramm aufgenommen werden sollen, und festlegen, ob Zentroide darin aufgenommen werden sollen.

Abbildung 3-11
Dialogfeld "Ladungsdiagramme"



Komponentenladungen anzeigen. Bei Auswahl dieser Option wird ein Diagramm der Komponentenladungen angezeigt.

Ladungsvariablen. Das Komponentenladungsdiagramm kann auf allen Variablen oder einer Teilmenge der Variablen basieren.

Zentroide einschließen. Bei Variablen mit mehrfach nominalem Skalierungsniveau sind keine Komponentenladungen vorhanden. Bei diesen Variablen können jedoch die Zentroide in das Diagramm aufgenommen werden. Sie können alle mehrfach nominalen Variablen verwenden oder eine Teilmenge auswählen.

Zusätzliche Funktionen beim Befehl CATPCA

Sie können die kategoriale Hauptkomponentenanalyse an Ihre Bedürfnisse anpassen, indem Sie Ihre Auswahl in ein Syntax-Fenster einfügen und die daraus resultierende Befehlssyntax für CATPCA bearbeiten. Mit der Befehlssyntax-Sprache verfügen Sie außerdem über folgende Möglichkeiten:

- Mit dem Unterbefehl SAVE können Sie Stammnamen für die transformierten Variablen, Objektwerte und Approximationen festlegen, wenn Sie diese in der Arbeitsdatei speichern.
- Mit dem Unterbefehl PLOT kann die Höchstlänge für Beschriftungen für jedes Diagramm getrennt festgelegt werden.
- Mit dem Unterbefehl PLOT kann eine separate Variablenliste für Residuen-Diagramme festgelegt werden.

Vollständige Informationen zur Syntax finden Sie in der *Command Syntax Reference*.

Kapitel
4

Nichtlineare kanonische Korrelationsanalyse (OVERALS)

Die nichtlineare kanonische Korrelationsanalyse entspricht der kategorialen kanonischen Korrelationsanalyse mit optimaler Skalierung. Ziel dieser Prozedur ist es, den Grad der gegenseitigen Ähnlichkeit von Sets kategorialer Variablen zu bestimmen. Die nichtlineare kanonische Korrelationsanalyse ist auch unter der Akronym OVERALS bekannt.

Die kanonische Standard-Korrelationsanalyse ist eine Erweiterung der mehrfachen Regression, wobei das zweite Set nicht nur eine Antwortvariable, sondern mehrere Antwortvariablen enthält. Ziel ist es, einen möglichst großen Teil der Varianz in der Beziehung zwischen zwei Sets von numerischen Variablen in einem Raum mit wenigen Dimensionen zu erklären. Zunächst werden die Variablen in jedem Set linear so kombiniert, dass die linearen Kombinationen maximal miteinander korrelieren. Nach diesen Kombinationen werden weitere lineare Kombinationen ermittelt, die nicht mit den vorherigen Kombinationen korrelieren und über die höchstmögliche Korrelation verfügen.

Der Einsatz der optimalen Skalierung erweitert die Standardanalyse in drei entscheidenden Punkten. Erstens ermöglicht OVERALS mehr als zwei Variablen-Sets. Zweitens können Variablen nominal, ordinal oder numerisch skaliert werden. Als Folge davon können nichtlineare Beziehungen zwischen Variablen analysiert werden. Drittens werden nicht die Korrelationen zwischen den Variablen-Sets maximiert, sondern die Sets werden mit einem unbekannten, durch die Objektwerte definierten Kompromiss-Set verglichen.

Beispiel. Mit der kategorialen kanonischen Korrelationsanalyse mit optimaler Skalierung kann die Beziehung zwischen zwei Variablen-Sets grafisch dargestellt werden, von denen das eine die Berufsgruppe und die Ausbildungsdauer in Jahren und das andere die Region des Wohnsitzes und das Geschlecht enthält. Die Analyse könnte beispielsweise ergeben, dass die Ausbildungsdauer und der Wohnsitz einen höheren Erklärungswert haben als die übrigen Variablen. Außerdem könnten Sie herausfinden, dass der Erklärungswert der Ausbildungsdauer in Jahren in der ersten Dimension am höchsten ist.

Statistiken und Diagramme. Häufigkeiten, Zentroide, Iterationsprotokoll, Objektwerte, Kategorienquantifikationen, Gewichtungen, Komponentenladungen, einfache und mehrfache Anpassung, Diagramme der Objektwerte, Diagramme der Kategorienkoordinaten, Diagramme der Komponentenladungen, Diagramme der Kategorienzentroide, Transformationsdiagramme.

Daten. Verwenden Sie zum Kodieren von kategorialen Variablen ganze Zahlen (nominales oder ordinales Skalierungsniveau). Um die Ausgabe zu minimieren, verwenden Sie zum Kodieren jeder Variablen aufeinander folgende ganze Zahlen, mit 1 beginnend. Variablen, die auf numerischem Niveau skaliert sind, dürfen nicht in aufeinander folgende ganze Zahlen umkodiert werden. Um die Ausgabe zu minimieren, subtrahieren Sie für jede auf numerischem Niveau skalierte Variable

den kleinsten beobachteten Wert von jedem Wert, und addieren Sie 1. Gebrochene Werte werden hinter der Dezimalstelle abgeschnitten.

Annahmen. Variablen können in ein oder mehr Sets klassifiziert werden. Variablen in der Analyse werden mehrfach nominal, einzeln nominal, ordinal oder numerisch skaliert. Die in der Prozedur verwendete maximale Anzahl von Dimensionen hängt vom Messniveau der optimalen Skalierung der Variablen ab. Wenn alle Variablen als ordinal, einzeln nominal oder numerisch angegeben sind, entspricht die maximale Anzahl von Dimensionen dem kleineren der folgenden beiden Werte: die Anzahl der Beobachtungen minus 1 oder die Gesamtanzahl der Variablen. Wenn jedoch nur zwei Variablen-Sets definiert sind, entspricht die maximale Anzahl von Dimensionen der Anzahl von Variablen im kleineren Set. Wenn einige Variablen mehrfach nominal sind, entspricht die maximale Anzahl von Dimensionen der Gesamtanzahl von mehrfach nominalen Kategorien plus der Anzahl von nicht mehrfach nominalen Variablen minus der Anzahl von mehrfach nominalen Variablen. Wenn die Analyse beispielsweise fünf Variablen umfasst, von denen eine mehrfach nominal mit vier Kategorien ist, ist die maximale Anzahl von Dimensionen gleich (4 + 4 – 1), also 7. Wenn Sie eine Zahl angeben, die größer als der Höchstwert ist, wird der Höchstwert verwendet.

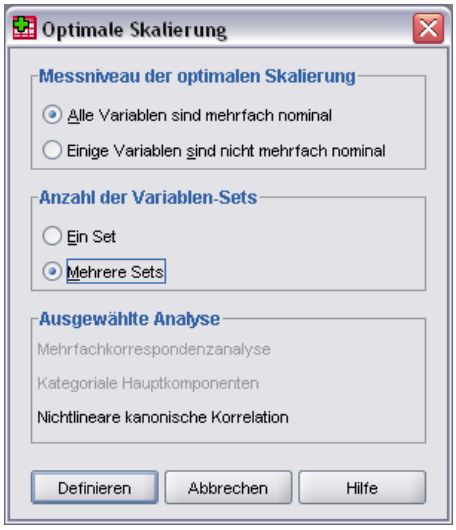
Verwandte Prozeduren. Wenn jedes Set eine Variable enthält, ist die nichtlineare kanonische Korrelationsanalyse äquivalent zur Hauptkomponentenanalyse mit optimaler Skalierung. Wenn jede dieser Variablen mehrfach nominal ist, entspricht die Analyse der Mehrfachkorrespondenzanalyse. Wenn zwei Variablen-Sets vorhanden sind und eines dieser Sets nur eine Variable enthält, ist die Analyse identisch mit der kategorialen Regression mit optimaler Skalierung.

So berechnen Sie eine nichtlineare kanonische Korrelationsanalyse:

▶ Wählen Sie die folgenden Befehle aus den Menüs aus:

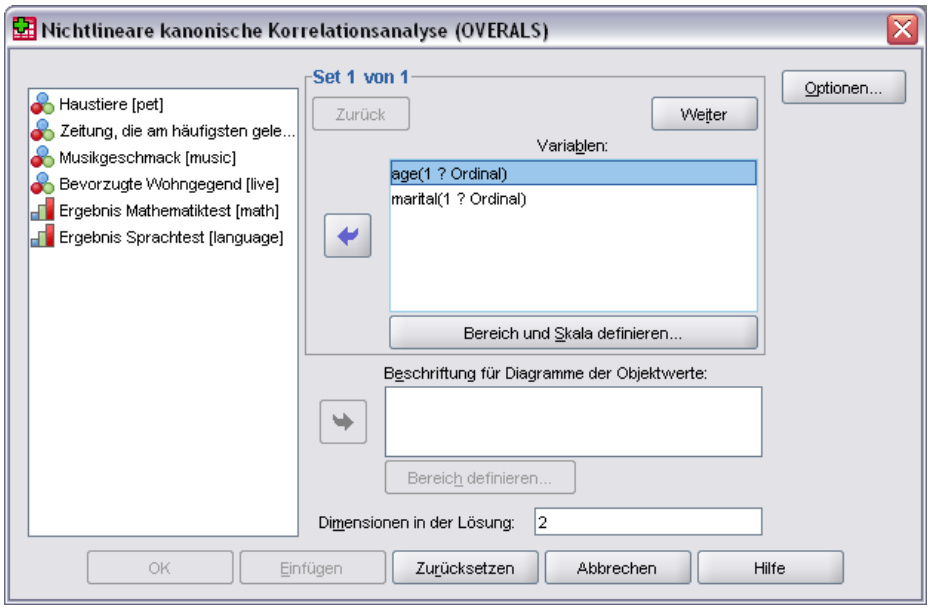
Analysieren
  Dimensionsreduktion
    Optimale Skalierung...

Abbildung 4-1
Dialogfeld "Optimale Skalierung"



- Wählen Sie entweder Alle Variablen sind mehrfach nominal oder Einige Variablen sind nicht mehrfach nominal aus.
- Wählen Sie Mehrere Sets.
- Klicken Sie auf Definieren.

Abbildung 4-2
Dialogfeld "Nichtlineare kanonische Korrelationsanalyse (OVERALS)"



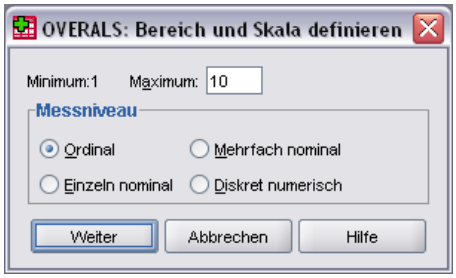
- Definieren Sie mindestens zwei Variablen-Sets. Wählen Sie die Variablen aus, die in das erste Set aufgenommen werden sollen. Um zum nächsten Set zu wechseln, klicken Sie auf Weiter und wählen Sie die Variablen aus, die in das zweite Set aufgenommen werden sollen. Auf Wunsch können Sie weitere Sets hinzufügen. Klicken Sie auf Zurück, um zum zuvor definierten Variablen-Set zurückzukehren.
- Legen Sie für jede Variable den Wertebereich und das Messniveau (Niveau der optimalen Skalierung) fest.
- Klicken Sie auf OK.
- Die folgenden Optionen sind verfügbar:
  - Sie können eine oder mehrere Variablen auswählen, um Punktbeschriftungen für Objektwertediagramme anzugeben. Jede Variable erzeugt ein separates Diagramm, dessen Punkte mit den Werten dieser Variablen beschriftet sind. Für jede dieser Variablen zur Diagrammbeschriftung muss ein Bereich festgelegt werden. In diesem Dialogfeld kann eine Variable nicht gleichzeitig in der Analyse und als Beschriftungsvariable verwendet werden. Wenn Sie das Objektwertediagramm mit einer in der Analyse verwendeten Variablen beschriften möchten, verwenden Sie die Funktion "Berechnen" (verfügbar im Menü "Transformieren"), um eine Kopie dieser Variablen zu erstellen. Verwenden Sie die

neue Variable, um das Diagramm zu beschriften. Wahlweise können Sie dazu auch die Befehlssyntax verwenden.

- Sie können die in der Lösung gewünschte Anzahl der Dimensionen angeben. Im allgemeinen sollten Sie nur so viele Dimensionen auswählen, wie zum Erklären des größten Teils der Variation notwendig ist. Wenn die Analyse mehr als zwei Dimensionen umfasst, werden dreidimensionale Diagramme der ersten drei Dimensionen erzeugt. Weitere Dimensionen können durch die Bearbeitung des Diagramms angezeigt werden.

Bereich und Skala definieren

Abbildung 4-3
Dialogfeld "Bereich und Skala definieren"



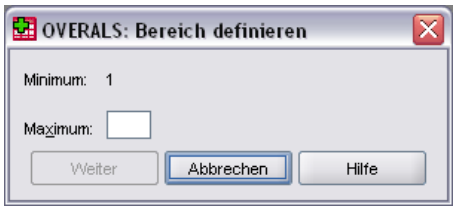
Sie müssen für jede Variable einen Bereich definieren. Der angegebene Höchstwert muss eine Ganzzahl sein. Gebrochene Datenwerte werden in der Analyse abgeschnitten. Ein Kategoriewert, der außerhalb des angegebenen Bereiches liegt, wird in der Analyse ignoriert. Um eine möglichst kleine Ausgabe zu erzielen, erstellen Sie mithilfe der Funktion "Automatisch umkodieren" (verfügbar im Menü "Transformieren") aufeinander folgende, mit 1 beginnende Kategorien für Variablen, die als nominale oder ordinale Variablen behandelt werden. Das Umkodieren in fortlaufende Ganzzahlen empfiehlt sich nicht für Variablen, die auf numerischem Niveau skaliert sind. Um bei Variablen, die als numerische Variablen behandelt werden, eine möglichst kleine Ausgabe zu erzielen, subtrahieren Sie für alle Variablen von jedem Wert den Minimalwert, und addieren Sie 1.

Sie müssen außerdem die für die Quantifikation der einzelnen Variablen verwendete Skalierung auswählen.

- **Ordinal.** Bei der beobachteten Variablen wird die Reihenfolge der Kategorien bei der quantifizierten Variablen beibehalten.
- **Einzeln nominal.** Bei der quantifizierten Variablen erhalten Objekte der gleichen Kategorie den gleichen Wert.
- **Mehrfach nominal.** Die Quantifikationen können bei den einzelnen Dimensionen unterschiedlich sein.
- **Diskret numerisch.** Die Kategorien werden als geordnet mit gleichen Abständen behandelt. Die Unterschiede zwischen den Kategorienummern und die Reihenfolge der Kategorien der beobachteten Variablen bleiben bei der quantifizierten Variablen erhalten.

Bereich definieren

Abbildung 4-4
Dialogfeld "Bereich definieren"



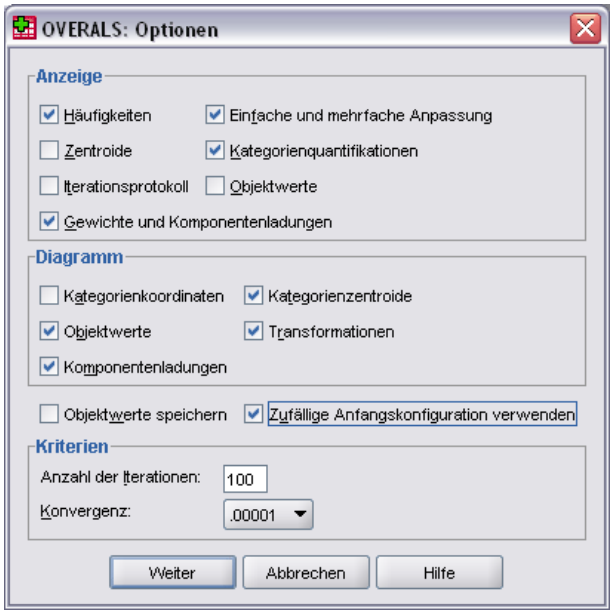
Sie müssen für jede Variable einen Bereich definieren. Der angegebene Höchstwert muss eine Ganzzahl sein. Gebrochene Datenwerte werden in der Analyse abgeschnitten. Ein Kategoriewert, der außerhalb des angegebenen Bereiches liegt, wird in der Analyse ignoriert. Um eine möglichst kleine Ausgabe zu erzielen, erstellen Sie mithilfe der Funktion "Automatisch umkodieren" (verfügbar im Menü "Transformieren") aufeinander folgende Kategorien, die mit 1 beginnen.

Ferner müssen Sie einen Bereich für jede Variable definieren, mit deren Hilfe die Diagramme der Objektwerte beschriftet werden. Beschriftungen für Kategorien mit Datenwerten, die außerhalb des für die Variable definierten Bereichs liegen, werden in den Diagrammen jedoch angezeigt.

Nichtlineare kanonische Korrelationsanalyse: Optionen

Das Dialogfeld "Optionen" ermöglicht Ihnen die Auswahl von optionalen Statistiken und Diagrammen, die Speicherung von Objektwerten als neue Variablen in der Arbeitsdatei, die Angabe von Iterations- und Konvergenzkriterien sowie die Angabe einer Ausgangskonfiguration für die Analyse.

Abbildung 4-5
Dialogfeld "Optionen"



Anzeigen. Die verfügbaren Statistiken umfassen: marginale Häufigkeiten, Zentroide, Iterationsprotokoll, Gewichte und Komponentenladungen, Kategorienquantifikationen, Objektwerte und einfache und mehrfache Anpassung.

- **Zentroide.** Kategorienquantifikationen sowie die prognostizierten und tatsächlichen Mittelwerte der Objektwerte für die in jedem Set enthaltenen Objekte (Fälle), die zur gleichen Kategorie wie die Variable gehören.
- **Gewichte und Komponentenladungen.** Der Regressionskoeffizient in jeder Dimension für jede quantifizierte Variable in einem Set, wobei für die Objektwerte eine Regression auf die quantifizierten Variablen durchgeführt wird, sowie die Projektion der quantifizierten Variablen im Objektraum. Liefert einen Hinweis über den Beitrag, den jede Variable zu der Dimension innerhalb eines Sets liefert.
- **Einfache und mehrfache Anpassung.** Maße für die Anpassungsgüte der einfachen und multiplen Kategorienkoordinaten bzw. Kategorienquantifizierungen hinsichtlich der Objekte.
- **Kategorienquantifikationen.** Zuordnung optimaler Skalenwerte zu den Kategorien einer Variablen.
- **Objektwerte.** Der optimale Wert (Score), der einem Objekt (Fall) in einer bestimmten Dimension zugewiesen wird.

Diagramm. Sie können Diagramme von Kategorienkoordinaten, Objektwerten, Komponentenladungen, Kategorienzentroiden und Transformationen erzeugen.

Objektwerte speichern. Sie können die Objektwerte als neue Variablen in der Arbeitsdatei speichern. Objektwerte werden für die im Hauptdialogfeld angegebene Anzahl von Dimensionen gespeichert.

Zufällige Anfangskonfiguration verwenden. Wenn einige oder alle Variablen einzeln nominal sind, sollte eine zufällige Ausgangskonfiguration verwendet werden. Wenn diese Option nicht ausgewählt wird, wird eine verschachtelte Ausgangskonfiguration verwendet.

Kriterien. Sie können die maximale Anzahl von Iterationen festlegen, die in der nichtlinearen kanonischen Korrelationsanalyse bei den Berechnungen durchlaufen werden. Außerdem können Sie einen Wert für das Konvergenzkriterium auswählen. Die Analyse wird abgebrochen, wenn die Differenz der Gesamtanpassung zwischen den letzten zwei Iterationen kleiner als der Konvergenzwert ist oder die Maximalzahl von Iterationen erreicht wurde.

Zusätzliche Funktionen beim Befehl OVERALS

Sie können die kategoriale Hauptkomponentenanalyse an Ihre Bedürfnisse anpassen, indem Sie Ihre Auswahl in ein Syntax-Fenster einfügen und die daraus resultierende Befehlssyntax für `OVERALS` bearbeiten. Mit der Befehlssyntax-Sprache verfügen Sie außerdem über folgende Möglichkeiten:

- Festlegen der in Diagrammen darzustellenden Dimensionspaare anstelle der Darstellung aller extrahierten Dimensionen (unter Verwendung des Schlüsselwortes `NDIM` für den Unterbefehl `PLOT`).
- Festlegen der Zeichenanzahl für Wertelabels zur Beschriftung der Punkte in den Diagrammen (mit dem Unterbefehl `PLOT`).

- Festlegen von mehr als fünf Variablen für die Diagramme der Objektwerte (mit dem Unterbefehl PLOT).
- Auswahl von in der Analyse verwendeten Variablen als Beschriftungsvariablen für die Diagramme der Objektwerte (mit dem Unterbefehl PLOT).
- Auswahl von Variablen als Punktbeschriftungen für die Diagramme der Quantifikationswerte (mit dem Unterbefehl PLOT).
- Festlegen der Anzahl von Fällen, die in die Analyse einbezogen werden sollen, falls Sie nicht alle Fälle in der Arbeitsdatei verwenden möchten (mit dem Unterbefehl NOBSERVATIONS).
- Festlegen der Stammnamen für Variablen, die durch Speichern der Objektwerte erstellt wurden (mit dem Unterbefehl SAVE).
- Festlegen der zu speichernden Anzahl an Dimensionen, sodass nicht alle extrahierten Dimensionen gespeichert werden (mit dem Unterbefehl SAVE).
- Schreiben von Kategoriequantifikationen in eine Matrixdatei (mit dem Unterbefehl MATRIX).
- Erzeugen von Diagrammen mit geringer Auflösung, die lesbarer sind als die üblichen hochauflösenden Diagramme (mit dem Befehl SET).
- Erstellen von Zentroid- und Transformationsdiagrammen für bestimmte Variablen (mit dem Unterbefehl PLOT).

Vollständige Informationen zur Syntax finden Sie in der *Command Syntax Reference*.

Kapitel
5

Korrespondenzanalyse

Ein Ziel der Korrespondenzanalyse besteht darin, die Beziehungen zwischen zwei nominalen Variablen in einer Korrespondenztabelle in einem flachdimensionierten Raum und gleichzeitig die Beziehungen zwischen den Kategorien für jede Variable zu beschreiben. Für jede Variable werden die Beziehungen zwischen den Kategorien durch die Distanzen zwischen den Kategoriepunkten in einem Diagramm dargestellt, wobei ähnliche Kategorien jeweils nahe beieinander liegen. Durch die Projizierung von Punkten für eine Variable auf dem Vektor vom Ursprung zu einem Kategoriepunkt einer anderen Variablen wird die Beziehung zwischen den Variablen beschrieben.

Bei einer Analyse der Kontingenztafeln werden häufig Zeilen- und Spaltenprofile untersucht und Tests auf Unabhängigkeit mithilfe der Chi-Quadrat-Statistik durchgeführt. Die Anzahl der Profile kann jedoch unter Umständen relativ groß sein, wobei durch die Chi-Quadrat-Statistik die Abhängigkeitsstruktur nicht erkennbar wird. Die Prozedur "Kreuztabelle" bietet verschiedene Zusammenhangsmaße und -tests, kann jedoch keine Beziehungen zwischen den Variablen darstellen.

Die Faktorenanalyse ist ein Standardverfahren zur Beschreibung von Beziehungen zwischen Variablen in einem flachdimensionierten Raum. Zur Faktorenanalyse werden jedoch Intervalldaten benötigt. Außerdem muss die Anzahl der Beobachtungen das Fünffache der Anzahl der Variablen betragen. Andererseits wird bei der Korrespondenzanalyse von nominalen Variablen ausgegangen, sodass die Beziehungen zwischen den Kategorien jeder Variablen sowie die Beziehungen zwischen den Variablen beschrieben werden können. Zudem kann die Korrespondenzanalyse zur Untersuchung einer beliebigen Tabelle mit positiven Korrespondenzmaßen verwendet werden.

Beispiel. Mithilfe der Korrespondenzanalyse kann die Beziehung zwischen einer Mitarbeitergruppe und den Rauchgewohnheiten grafisch dargestellt werden. Es könnte sich herausstellen, dass die Rauchgewohnheiten bei Angestellten im mittleren Management und beim Sekratariatspersonal jeweils unterschiedlich, bei Sekretariatsmitarbeitern im Vergleich zu Angestellten im gehobenen Management jedoch identisch sind. Als Ergebnis könnte sich ebenfalls die Erkenntnis herausstellen, dass starkes Rauchen mit Angestellten des mittleren Managements und gelegentliches Rauchen mit Sekretärinnen in Verbindung gebracht werden kann.

Statistiken und Diagramme. Korrespondenzmaße, Zeilen- und Spaltenprofile, Singulärwerte, Zeilen- und Spaltenwerte, Trägheit, Masse, Konfidenzstatistik für Zeilen- und Spaltenwerte, Konfidenzstatistik für Singulärwerte, Transformationsdiagramme, Zeilenpunktdiagramme, Spaltenpunktdiagramme und Biplots.

Daten. Die kategorialen Variablen, die analysiert werden sollen, werden nominal skaliert. Verwenden Sie für aggregierte Daten oder für ein Korrespondenzmaß (außer Häufigkeiten) eine Gewichtungsvariable mit positiven Ähnlichkeitswerten. Sie können für Tabellendaten auch die Syntax zum Lesen der Tabelle verwenden.

Annahmen. Die maximale Anzahl der in der Prozedur verwendeten Dimensionen ist von der Zahl der aktiven Zeilen- und Spaltenkategorien und der Anzahl der Gleichheitsbedingungen abhängig. Wenn keine Gleichheitsbedingungen verwendet werden und alle Kategorien aktiv sind, ist die maximale Dimensionalität um eins kleiner als die Anzahl der Kategorien in der Variablen mit den wenigsten Kategorien. Wenn beispielsweise eine Variable über fünf und eine andere über vier Kategorien verfügt, beträgt die maximale Anzahl an Dimensionen drei. Zusätzliche Kategorien sind nicht aktiv. Wenn beispielsweise eine Variable über fünf Kategorien (davon zwei Zusatzkategorien) und eine andere über vier Kategorien verfügt, beträgt die maximale Anzahl an Dimensionen zwei. Behandeln Sie alle Kategoriegruppen, die auf Gleichheit beschränkt sind, als eine Kategorie. Wenn beispielsweise eine Variable über fünf Kategorien verfügt, von denen drei auf Gleichheit beschränkt sein müssen, muss diese Variable bei der Bestimmung der maximalen Dimensionalität so behandelt werden, als ob sie drei Kategorien besitzen würde. Für zwei der Kategorien gelten keine Nebenbedingungen, und die dritte Kategorie entspricht den drei Kategorien mit der Gleichheitsbeschränkung. Wenn Sie für die Anzahl der Dimensionen eine über dem Maximalwert liegende Zahl angeben, wird in jedem Fall der Maximalwert verwendet.

Verwandte Prozeduren. Falls Sie mit mehr als zwei Variablen arbeiten, verwenden Sie die Mehrfachkorrespondenzanalyse. Wenn die Variablen ordinal skaliert werden sollen, verwenden Sie die kategoriale Hauptkomponentenanalyse.

So lassen Sie eine Korrespondenzanalyse berechnen:

- Wählen Sie die folgenden Befehle aus den Menüs aus:

  Analysieren
  Dimensionsreduktion
  Korrespondenzanalyse...

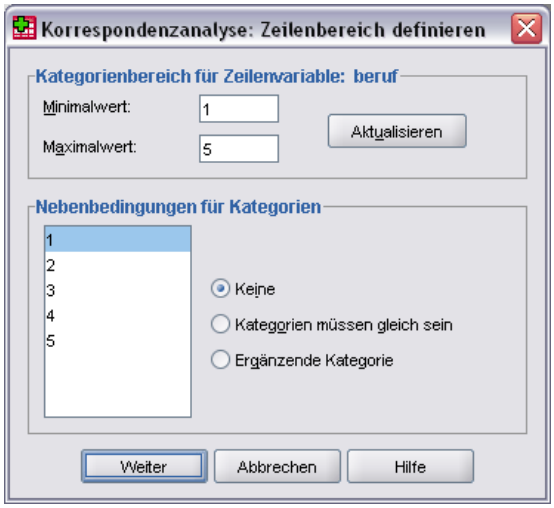
Abbildung 5-1
Dialogfeld "Korrespondenzanalyse"

- Wählen Sie eine Zeilenvariable aus.
- Wählen Sie eine Spaltenvariable aus.
- Definieren Sie die Bereiche für die Variablen.
- Klicken Sie auf OK.

Definieren des Zeilenbereichs bei der Korrespondenzanalyse

Sie müssen einen Bereich für die Zeilenvariable angeben. Die angegebenen Minimal- und Maximalwerte müssen ganzzahlig sein. Gebrochene Datenwerte werden in der Analyse abgeschnitten. Ein Kategoriewert, der außerhalb des angegebenen Bereiches liegt, wird in der Analyse ignoriert.

Abbildung 5-2
Dialogfeld "Zeilenbereich definieren"



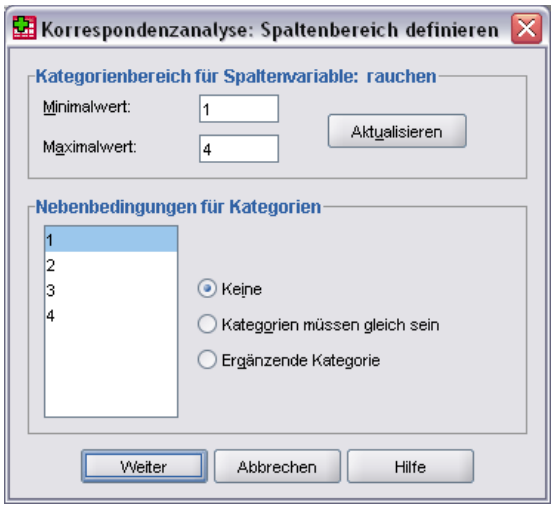
Alle Kategorien sind zunächst aktiv und nicht durch Nebenbedingungen eingeschänkt. Sie können festlegen, dass bestimmte Zeilenkategorien anderen Zeilenkategorien entsprechen müssen. Außerdem können Sie eine Zeilenkategorie als ergänzende Kategorie definieren.

- **Kategorien müssen gleich sein.** Die Kategorien müssen gleiche Werte haben. Verwenden Sie Gleichheitsbeschränkungen, wenn die erhaltene Ordnung der Kategorien nicht erwünscht oder nicht intuitiv ist. Die maximale Anzahl der Zeilenkategorien, für die Gleichheitsbeschränkungen festgelegt werden können, entspricht der Gesamtzahl aktiver Zeilenkategorien minus 1. Zum Definieren anderer Gleichheitsbeschränkungen für Kategoriegruppen können Sie die Syntax verwenden. Beispielsweise können Sie mithilfe der Syntax festlegen, dass die Kategorien 1 und 2 sowie die Kategorien 3 und 4 gleich sein müssen.
- **Ergänzende Kategorie.** Ergänzende Kategorien haben keine Auswirkungen auf die Analyse, werden jedoch in dem durch die aktiven Kategorien definierten Raum dargestellt. Ergänzende Kategorien spielen keine Rolle bei der Definition von Dimensionen. Die maximale Anzahl der ergänzenden Zeilenkategorien entspricht der Gesamtanzahl der Zeilenkategorien minus 2.

Definieren des Spaltenbereichs in der Korrespondenzanalyse

Sie müssen einen Bereich für die Spaltenvariable angeben. Die angegebenen Minimal- und Maximalwerte müssen ganzzahlig sein. Gebrochene Datenwerte werden in der Analyse abgeschnitten. Ein Kategoriewert, der außerhalb des angegebenen Bereiches liegt, wird in der Analyse ignoriert.

Abbildung 5-3
Dialogfeld "Spaltenbereich definieren"



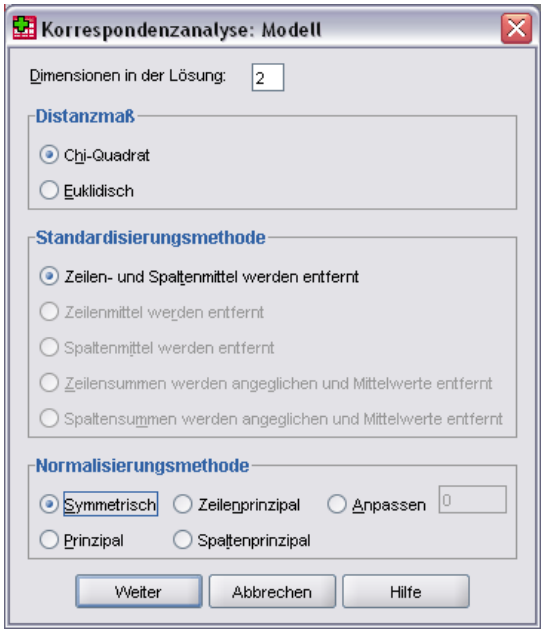
Alle Kategorien sind zunächst aktiv und nicht durch Nebenbedingungen eingeschänkt. Sie können festlegen, dass bestimmte Spaltenkategorien anderen Spaltenkategorien entsprechen müssen. Außerdem können Sie eine Spaltenkategorien als ergänzende Kategorie definieren.

- **Kategorien müssen gleich sein.** Die Kategorien müssen gleiche Werte haben. Verwenden Sie Gleichheitsbeschränkungen, wenn die erhaltene Ordnung der Kategorien nicht erwünscht oder nicht intuitiv ist. Die maximale Anzahl der Spaltenkategorien, für die Gleichheitsbeschränkungen festgelegt werden können, entspricht der Gesamtzahl aktiver Spaltenkategorien minus 1. Zum Definieren anderer Gleichheitsbeschränkungen für Kategoriegruppen können Sie die Syntax verwenden. Beispielsweise können Sie mithilfe der Syntax festlegen, dass die Kategorien 1 und 2 sowie die Kategorien 3 und 4 gleich sein müssen.
- **Ergänzende Kategorie.** Ergänzende Kategorien haben keine Auswirkungen auf die Analyse, werden jedoch in dem durch die aktiven Kategorien definierten Raum dargestellt. Ergänzende Kategorien spielen keine Rolle bei der Definition von Dimensionen. Die maximale Anzahl der ergänzenden Spaltenkategorien entspricht der Gesamtanzahl der Spaltenkategorien minus 2.

Korrespondenzanalyse: Modell

Im Dialogfeld "Modell" können Sie die Anzahl der Dimensionen, die Distanzmaße sowie die Standardisierungsmethode und die Normalisierungsmethode angeben.

Abbildung 5-4
Dialogfeld "Modell"



Dimensionen in der Lösung. Legen Sie die Anzahl der Dimensionen fest. Im allgemeinen sollten Sie nur so viele Dimensionen auswählen, wie zum Erklären des größten Teils der Variation notwendig ist. Die maximal zulässige Dimensionszahl hängt von der Anzahl der in der Analyse und in den Gleichheitsbeschränkungen verwendeten aktiven Kategorien ab. Die maximale Anzahl an Dimensionen ist jeweils der kleinere von folgenden Werten:

- Die Anzahl der aktiven Zeilenkategorien minus der Anzahl der Zeilenkategorien, die gleich sein müssen, plus der Anzahl der eingeschränkten Zeilenkategoriengruppen.
- Die Anzahl der aktiven Spaltenkategorien minus der Anzahl der Spaltenkategorien, die gleich sein müssen, plus der Anzahl der eingeschränkten Spaltenkategoriengruppen.

Distanzmaß. Sie können das Distanzmaß für Zeilen und Spalten der Korrespondenztabelle auswählen. Wählen Sie eine der folgenden Optionen:

- **Chi-Quadrat.** Verwendet eine gewichtete Profildistanz, wobei die Gewichtung der Masse der Zeilen oder Spalten entspricht. Dieses Maß ist für die Standard-Korrespondenzanalyse erforderlich.
- **Euklidisch.** Verwendet die Quadratwurzel der Summe aus den quadrierten Differenzen von Zeilen- und Spaltenpaaren.

Standardisierungsmethode. Wählen Sie eine der folgenden Optionen:

- **Zeilen- und Spaltenmittel werden entfernt.** Sowohl Zeilen als auch Spalten werden zentriert. Diese Methode ist für die Standard-Korrespondenzanalyse erforderlich.
- **Zeilenmittel werden entfernt.** Nur die Zeilen werden zentriert.
- **Spaltenmittel werden entfernt.** Nur die Spalten werden zentriert.

- **Zeilensummen werden angeglichen und Mittelwerte entfernt.** Vor dem Zentrieren der Zeilen werden die Zeilenränder angeglichen.
- **Spaltensummen werden angeglichen und Mittelwerte entfernt.** Vor dem Zentrieren der Spalten werden die Spaltenränder angeglichen.

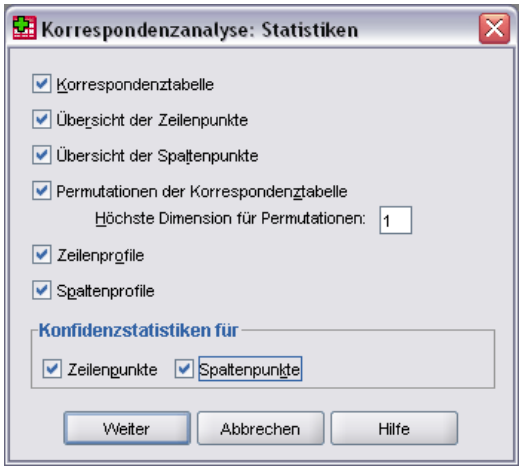
Normalisierungsmethode. Wählen Sie eine der folgenden Optionen:

- **Symmetrisch.** Für jede Dimension sind die Zeilenwerte der gewichtete Durchschnitt der Spaltenwerte geteilt durch den entsprechenden Singulärwert, und die Spaltenwerte sind der gewichtete Durchschnitt der Zeilenwerte geteilt durch den entsprechenden Singulärwert. Verwenden Sie diese Methode, wenn Sie die Unterschiede oder Ähnlichkeiten zwischen Kategorien der beiden Variablen untersuchen möchten.
- **Prinzipal**. Die Distanzen zwischen Zeilenpunkten und Spaltenpunkten sind Approximationen der Distanzen in der Korrespondenztabelle entsprechend dem gewählten Distanzmaß. Verwenden Sie diese Methode, wenn Sie die Unterschiede zwischen Kategorien einer oder in beiden Variablen anstelle der Unterschiede zwischen den beiden Variablen untersuchen möchten.
- **Zeilenprinzipal.** Die Distanzen zwischen Zeilenpunkten sind Approximationen der Distanzen in der Korrespondenztabelle entsprechend dem gewählten Distanzmaß. Die Zeilenwerte entsprechen dem gewichteten Durchschnitt der Spaltenwerte. Verwenden Sie diese Methode, wenn Sie die Unterschiede oder Ähnlichkeiten zwischen Kategorien der Zeilenvariable untersuchen möchten.
- **Spaltenprinzipal.** Die Distanzen zwischen Spaltenpunkten sind Approximationen der Distanzen in der Korrespondenztabelle entsprechend dem gewählten Distanzmaß. Die Spaltenwerte entsprechen dem gewichteten Durchschnitt der Zeilenwerte. Verwenden Sie diese Methode, wenn Sie die Unterschiede oder Ähnlichkeiten zwischen Kategorien der Spaltenvariablen untersuchen möchten.
- **Anpassen.** Sie müssen einen Wert zwischen –1 und 1 angeben. Der Wert –1 entspricht dem Spaltenprinzipal. Der Wert 1 entspricht dem Zeilenprinzipal. Der Wert 0 entspricht der Einstellung “Symmetrisch”. Bei allen übrigen Werten wird die Trägheit in unterschiedlichem Maße über Zeilen- und Spaltenwerte gestreut. Mit dieser Methode können Sie Biplots genau an Ihre Bedürfnisse anpassen.

Korrespondenzanalyse: Statistiken

Im Dialogfeld “Statistiken” können Sie festlegen, welche numerische Ausgabe erzeugt werden soll.

Abbildung 5-5
Dialogfeld "Statistiken"



Korrespondenztabelle. Eine Kreuztabelle für die Eingabevariable mit Randhäufigkeiten der Zeilen und Spalten.

Übersicht der Zeilenpunkte. Für jede Zeilenkategorie werden die Werte, die Masse, die Trägheit, der Anteil der Dimension an der Trägheit und der Anteil der Dimension an der Trägheit des Punkts angezeigt.

Übersicht der Spaltenpunkte. Für jede Spaltenkategorie werden die Werte, die Masse, die Trägheit, der Anteil der Dimension an der Trägheit und der Anteil der Dimension an der Trägheit des Punkts angezeigt.

Zeilenprofile. Für jede Zeilenkategorie wird die Verteilung über die Kategorien der Spaltenvariablen angezeigt.

Spaltenprofile. Für jede Spaltenkategorie wird die Verteilung über die Kategorien der Zeilenvariablen angezeigt.

Permutationen der Korrespondenztabelle. Die Korrespondenztabelle wird so umstrukturiert, dass die Zeilen und Spalten entsprechend den Werten der ersten Dimension in aufsteigender Reihenfolge angezeigt werden. Optional können Sie auch eine maximale Dimensionszahl angeben, für die Permutationstabellen erzeugt werden sollen. Dadurch wird für jede Dimension von 1 bis zur angegebenen Zahl eine Permutationstabelle erzeugt.

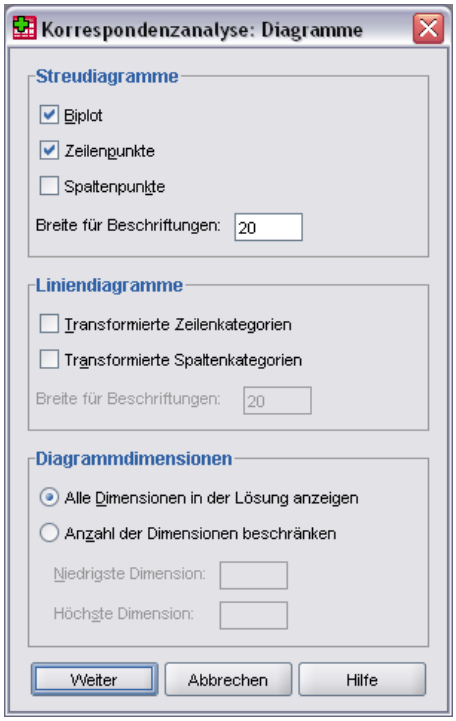
Konfidenzstatistiken für Zeilenpunkte. Bezieht die Standardabweichung und Korrelationen für alle nicht ergänzenden Zeilenpunkte mit ein.

Konfidenzstatistiken für Spaltenpunkte. Bezieht die Standardabweichung und Korrelationen für alle nicht ergänzenden Spaltenpunkte mit ein.

Korrespondenzanalyse: Diagramme

Im Dialogfeld "Diagramme" können Sie festlegen, welche Diagramme erzeugt werden sollen.

Abbildung 5-6
Dialogfeld "Diagramme"



Streudiagramme. Erzeugt eine Matrix mit allen paarweisen Diagrammen der Dimensionen. Folgende Streudiagramme sind verfügbar:

- **Biplot.** Erzeugt eine Matrix mit verbundenen Diagrammen der Zeilen- und Spaltenpunkte. Bei Wahl der Normalisierungsmethode "Prinzipal" ist die Option "Biplot" nicht verfügbar.
- **Zeilenpunkte.** Erzeugt eine Matrix mit Diagrammen der Zeilenpunkte.
- **Spaltenpunkte.** Erzeugt eine Matrix mit Diagrammen der Spaltenpunkte.

Optional können Sie angeben, wie viele Zeichen in Wertelabels zur Beschriftung der Punkte verwendet werden sollen. Dieser Wert muss eine nichtnegative Ganzzahl sein, die kleiner oder gleich 20 ist.

Liniendiagramme. Erzeugt ein Diagramm für jede Dimension der gewählten Variable. Folgende Liniendiagramme sind verfügbar:

- **Transformierte Zeilenkategorien.** Erzeugt ein Diagramm unter Verwendung der ursprünglichen Zeilenkategoriewerte und der entsprechenden Zeilenwerte.
- **Transformierte Spaltenkategorien.** Erzeugt ein Diagramm unter Verwendung der ursprünglichen Spaltenkategoriewerte und der entsprechenden Spaltenwerte.

Optional können Sie angeben, wie viele Zeichen in Wertelabels zur Beschriftung der Kategorienachse verwendet werden sollen. Dieser Wert muss eine nichtnegative Ganzzahl sein, die kleiner oder gleich 20 ist.

Diagrammdimensionen. Hiermit legen Sie fest, wie die Statistiken in der Ausgabe angezeigt werden.

- **Alle Dimensionen in der Lösung anzeigen.** Alle Dimensionen in der Lösung werden in einer Streudiagramm-Matrix angezeigt.
- **Anzahl der Dimensionen beschränken.** Die angezeigten Dimensionen werden auf geplottete Paare beschränkt. Wenn Sie die Dimensionen beschränken, müssen Sie die niedrigste und die höchste Dimension auswählen, die jeweils geplottet werden soll. Die niedrigste Dimension muss im Bereich von 1 bis zur Anzahl der Dimensionen in der Lösung minus 1 liegen und wird gegen höhere Dimensionen geplottet. Der Wert für die höchste Dimension muss im Bereich von 2 bis zur Anzahl der Dimensionen in der Lösung liegen und gibt die höchste Dimension an, die beim Plotten der Dimensionspaare verwendet wird. Diese Spezifikation gilt für alle angeforderten multidimensionalen Plots.

Zusätzliche Funktionen beim Befehl CORRESPONDENCE

Sie können die Korrespondenzanalyse an Ihre Bedürfnisse anpassen, wenn Sie Ihre Auswahl in ein Syntax-Fenster einfügen und die resultierende Befehlssyntax für den Befehl `CORRESPONDENCE` bearbeiten. Mit der Befehlssyntax-Sprache verfügen Sie außerdem über folgende Möglichkeiten:

- Angeben der Tabellendaten als Eingabedaten anstelle der Verwendung fallweise gebundener Daten (Unterbefehl `TABLE = ALL`).
- Angeben der Anzahl von Zeichen für Wertelabels zur Beschriftung von Punkten für jeden Typ der Streudiagramm- oder Biplot-Matrix (mit dem Unterbefehl `PLOT`).
- Angeben der Anzahl von Zeichen für Wertelabels zur Beschriftung von Punkten für jeden Typ von Liniendiagramm (mit dem Unterbefehl `PLOT`).
- Schreiben einer Matrix mit Zeilen- und Spaltenscores in eine Matrixdatendatei (mit dem Unterbefehl `OUTFILE`).
- Schreiben einer Matrix der Konfidenzstatistiken (Varianzen und Kovarianzen) für die Singulärwerte und die Scores in eine Matrixdatendatei (mit dem Unterbefehl `OUTFILE`).
- Festlegen der Gleichheit für mehrere Kategoriegruppen (mit dem Unterbefehl `EQUAL`).

Vollständige Informationen zur Syntax finden Sie in der *Command Syntax Reference*.

Kapitel
6

Mehrfachkorrespondenzanalyse

Die Mehrfachkorrespondenzanalyse quantifiziert nominale (kategoriale) Daten durch Zuweisung numerischer Werte zu den Fällen (Objekten) und Kategorien, sodass Objekte in derselben Kategorie eng beieinander und Objekte in verschiedenen Kategorien weit voneinander entfernt liegen. Jedes Objekt liegt so nah wie möglich an den Kategoriepunkten der Kategorien, die für dieses Objekt gelten. So unterteilen die Kategorien die Objekte in homogene Untergruppen. Variablen werden als homogen betrachtet, wenn sie Objekte, die in denselben Kategorien vorliegen, auch in dieselben Untergruppen einordnen.

Beispiel. Mithilfe der Mehrfachkorrespondenzanalyse könnte die Beziehung zwischen Berufsgruppe, Minderheit und Geschlecht grafisch dargestellt werden. Unter Umständen stellt sich dabei heraus, dass Minderheit und Geschlecht im Gegensatz zur Berufsgruppe klare Unterscheidungsmerkmale für Personen darstellen. Außerdem könnte sich ergeben, dass die Kategorien "Latino" und "Afro-Amerikaner" einander sehr ähnlich sind.

Statistiken und Diagramme. Objektwerte Diskriminationsmaße, Iterationsprotokoll, Korrelationen der ursprünglichen und transformierten Variablen, Kategorienquantifikationen, deskriptive Statistiken, Objektpunktdiagramme, Biplots, Kategoriendiagramme, verbundene Kategoriendiagramme, Transformationsdiagramme und Diskriminationsmaß-Protokolle.

Daten. Werte von String-Variablen werden immer in aufsteigender alphanumerischer Reihenfolge in positive Ganzzahlen umgewandelt. Benutzerdefinierte und systemdefinierte fehlende Werte sowie Werte kleiner als 1 werden als fehlend betrachtet. Sie können die Daten umkodieren oder eine Konstante zu Variablen mit Werten kleiner als 1 addieren, um die Werte als nichtfehlend zu deklarieren.

Annahmen. Alle Variablen weisen ein mehrfach nominales Skalierungsniveau auf. Die Daten müssen mindestens drei gültige Fälle enthalten. Die Analyse basiert auf Daten, die aus positiven ganzen Zahlen bestehen. Mit der Option zur Diskretisierung werden Variablen mit Dezimalwerten automatisch kategorisiert, indem die Werte in nahezu normalverteilte Kategorien gruppiert werden, und Werte von String-Variablen werden automatisch in positive Ganzzahlen umgewandelt. Sie können andere Diskretisierungsmethoden bestimmen.

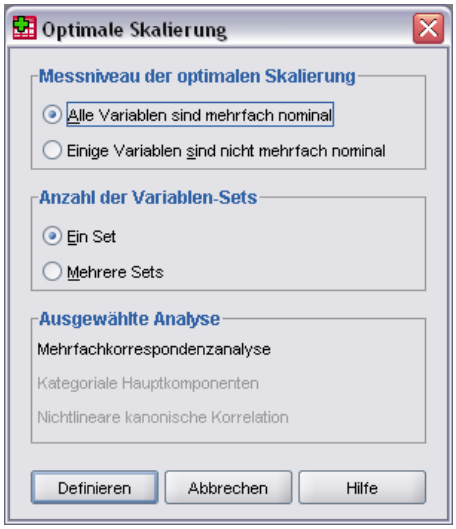
Verwandte Prozeduren. Bei zwei Variablen entspricht die Mehrfachkorrespondenzanalyse der Korrespondenzanalyse. Wenn Sie annehmen, dass die Variablen ordinale oder numerische Eigenschaften besitzen, sollte die kategoriale Hauptkomponentenanalyse verwendet werden. Wenn Variablen-Sets untersucht werden sollen, verwenden Sie die nichtlineare kanonische Korrelationsanalyse.

So lassen Sie eine Mehrfachkorrespondenzanalyse berechnen:

- Wählen Sie die folgenden Befehle aus den Menüs aus:

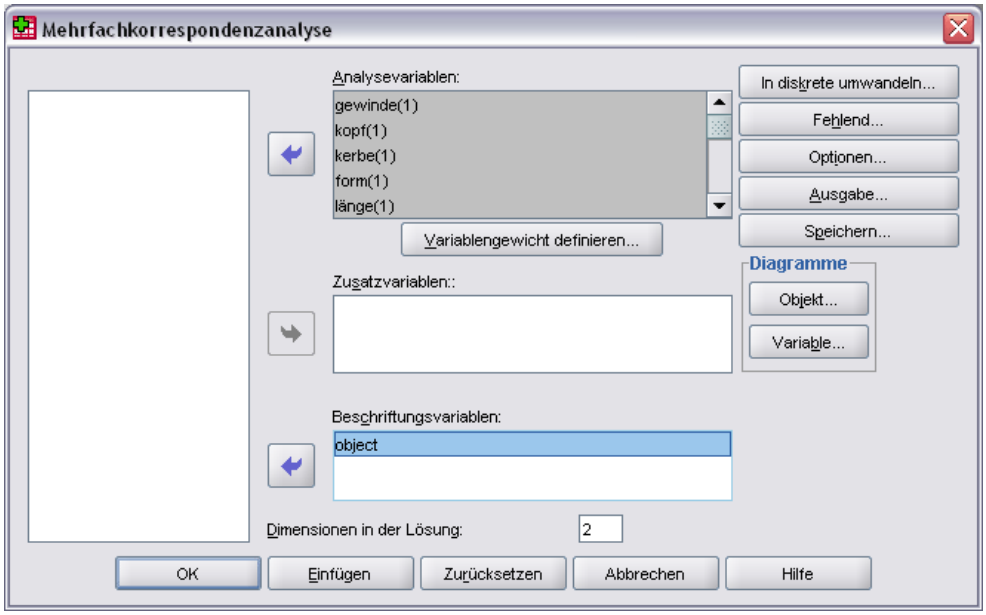
  Analysieren
  Dimensionsreduktion
  Optimale Skalierung...

Abbildung 6-1
Dialogfeld "Optimale Skalierung"



- Wählen Sie die Option Alle Variablen sind mehrfach nominal aus.
- Wählen Sie die Option Ein Set.
- Klicken Sie auf Definieren.

Abbildung 6-2
Dialogfeld "Mehrfachkorrespondenzanalyse"



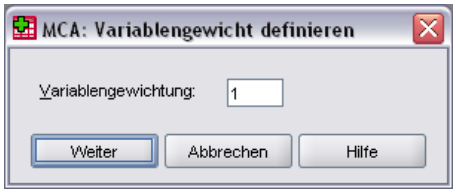
- Wählen Sie mindestens zwei Analysevariablen aus und geben Sie die Anzahl der Dimensionen in der Lösung an.
- Klicken Sie auf OK.

Sie können außerdem Beschriftungsvariablen für die Diagramme und Zusatzvariablen bestimmen, die in die gefundene Lösung eingepaßt werden.

Definition des Variablengewichts bei der Mehrfachkorrespondenzanalyse

Sie können das Gweicht für die Analysevariablen festlegen.

Abbildung 6-3
Dialogfeld "Variablengewicht definieren"

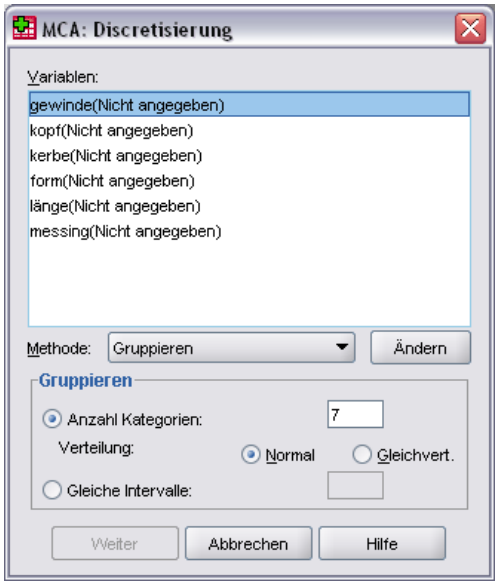


Variablengewichtung. Sie können für jede Variable eine Gewichtung festlegen. Bei dem angegebenen Wert muss es sich um eine positive ganze Zahl handeln. Der Standardwert ist 1.

Mehrfachkorrespondenzanalyse: Diskretisierung

Im Dialogfeld “Kategoriale Hauptkomponenten: Diskretisierung” können Sie eine Methode zum Umkodieren der Variablen festlegen. Sofern nicht anders angegeben, werden Variablen mit Dezimalwerten in sieben Kategorien gruppiert, die nahezu normalverteilt sind. (Wenn die Variable weniger als sieben unterschiedliche Werte aufweist, entspricht die Anzahl der Kategorien der Anzahl der unterschiedlichen Werte.) Werte von String-Variablen werden immer in positive ganze Zahlen umgewandelt, indem in aufsteigender alphanumerischer Reihenfolge Kategorienummern zugewiesen werden. Die Diskretisierung von String-Variablen erfolgt anhand dieser Ganzzahlen. Andere Variablen werden in der Standardeinstellung nicht modifiziert. Die diskretisierten Variablen werden dann in der Analyse verwendet.

Abbildung 6-4
Dialogfeld “Diskretisierung”



Methode. Verfügbar sind die Optionen “Gruppieren”, “Einstufen” oder “Multiplizieren”.

- **Gruppieren.** Hierbei werden die Daten in eine bestimmte Anzahl von Kategorien oder nach Intervall umkodiert.
- **Einstufen.** Die Variable wird durch Bilden einer Rangfolge diskretisiert.
- **Multiplizieren.** Die aktuellen Werte der Variablen werden standardisiert, mit 10 multipliziert und gerundet. Anschließend wird eine Konstante zu den Werten addiert, sodass der niedrigste diskretisierte Wert 1 beträgt.

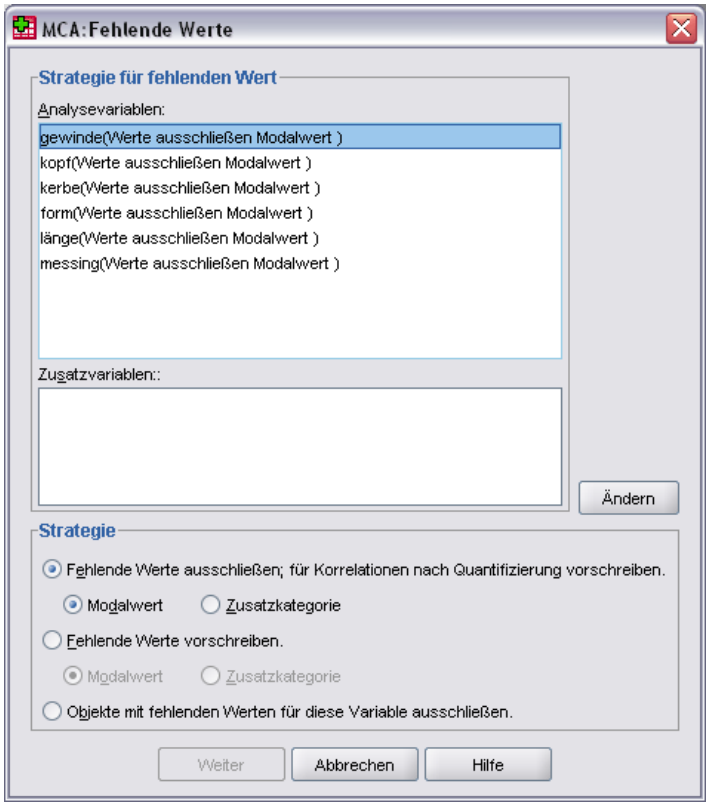
Gruppieren. Folgende Optionen sind bei der Diskretisierung von Variablen durch Gruppierung verfügbar:

- **Anzahl Kategorien.** Geben Sie eine Anzahl von Kategorien an. Legen Sie außerdem fest, ob die Werte der Variablen über die Kategorien ungefähr normal- oder gleichverteilt werden sollen.
- **Gleiche Intervalle.** Die Variablen werden in Kategorien umkodiert, die durch diese gleich großen Intervalle definiert werden. Sie müssen die Größe der Intervalle angeben.

Mehrfachkorrespondenzanalyse: Fehlende Werte

Im Dialogfeld “Kategoriale Hauptkomponenten: Fehlende Werte” können Sie die Strategie für die Verarbeitung von fehlenden Werten in den Analyse- und Zusatzvariablen festlegen.

Abbildung 6-5
Dialogfeld “Fehlende Werte”



Strategie für fehlenden Wert. Fehlende Werte können ausgeschlossen (Passivbehandlung) oder vorgeschrieben (Aktivbehandlung) werden oder Objekte mit fehlenden Werten können ausgeschlossen werden (listenweiser Fallausschluss).

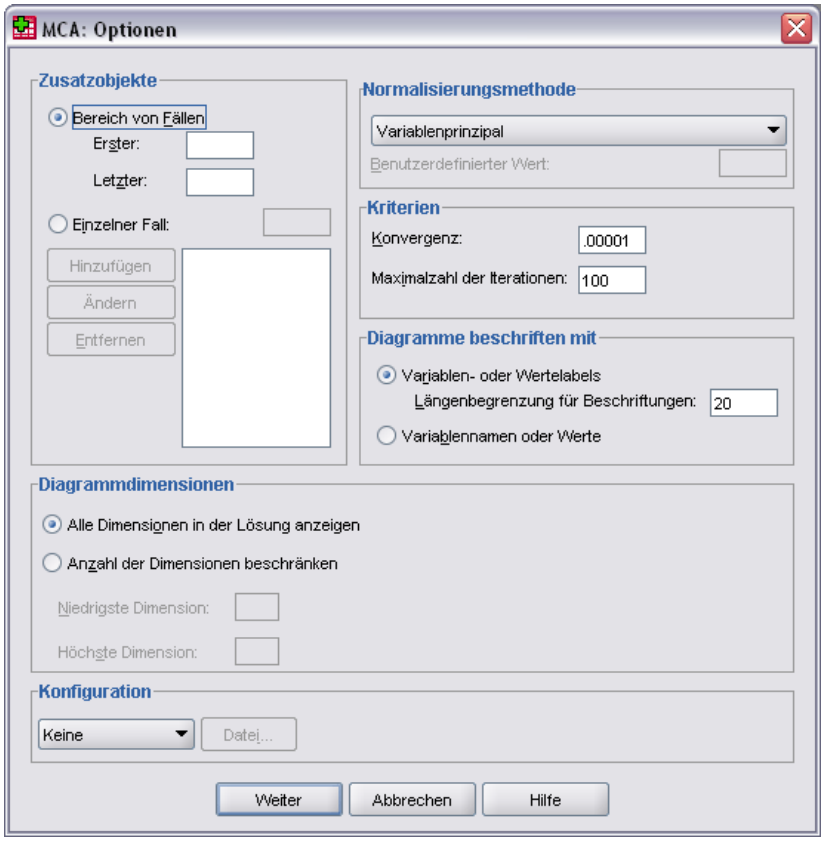
- **Fehlende Werte ausschließen; für Korrelationen nach Quantifizierung vorschreiben.** Objekte mit fehlenden Werten bei der ausgewählten Variablen tragen nicht zur Analyse dieser Variablen bei. Wenn alle Variablen passiv behandelt werden, werden Objekte mit fehlenden Werten für alle Variablen als Zusatzobjekte behandelt. Wenn im Dialogfeld “Kategoriale Hauptkomponenten: Ausgabe” Korrelationen ausgewählt sind, werden fehlende Werte nach der Analyse anhand der häufigsten Variablenkategorie (Modalwert) für die Korrelationen der ursprünglichen Variablengesetzt. Sie können die Methode auswählen, nach der Werte für die Korrelationen der optimal skalierten Variablen gesetzt werden. Wenn Sie Modalwert auswählen, werden fehlende Werte durch den Modalwert der optimal skalierten Variablen ersetzt. Wenn Sie Zusatzkategorie auswählen, werden fehlende Werte durch die Quantifikation einer zusätzlichen Kategorie ersetzt. Dies bedeutet aber auch, dass Objekte mit einem fehlenden Wert für diese Variable als zur gleichen Kategorie (der zusätzlichen) zugehörig betrachtet werden.

- **Fehlende Werte ersetzen.** Fehlende Werte werden bei Objekten mit fehlenden Werten für die ausgewählte Variable gesetzt. Sie können die Methode auswählen, nach der die Werte gesetzt werden: Wählen Sie Modalwert, wenn fehlende Werte durch die häufigste Kategorie ersetzt werden sollen. Falls mehrere Modalwerte vorhanden sind, wird der Modalwert mit der kleinsten Kategorienummer verwendet. Wenn Sie Zusatzkategorie auswählen, werden fehlende Werte durch die gleiche Quantifikation einer zusätzlichen Kategorie ersetzt. Dies bedeutet aber auch, dass Objekte mit einem fehlenden Wert für diese Variable als zur gleichen Kategorie (der zusätzlichen) zugehörig betrachtet werden.
- **Objekte mit fehlenden Werten für diese Variable ausschließen.** Objekte mit fehlenden Werten für die ausgewählte Variable werden aus der Analyse ausgeschlossen. Diese Strategie ist für Zusatzvariablen nicht verfügbar.

Mehrfachkorrespondenzanalyse: Optionen

Im Dialogfeld "Kategoriale Hauptkomponenten: Optionen" können Sie die anfängliche Konfiguration auswählen, Kriterien für die Iterationen und die Konvergenz festlegen, eine Normalisierungsmethode auswählen, die Methode zum Beschriften der Diagramme auswählen und Zusatzobjekte angeben.

Abbildung 6-6
Dialogfeld "Optionen"



Zusatzobjekte. Geben Sie die Fallnummer des Objekts (oder die erste und letzte Fallnummer einer Reihe von Objekten) an, die als Zusatzobjekte deklariert werden sollen, und klicken Sie dann auf Hinzufügen. Wiederholen Sie diesen Vorgang, bis Sie alle Zusatzobjekte festgelegt haben. Wenn ein Objekt als Zusatzobjekt deklariert ist, werden Fallgewichtungen für das Objekt ignoriert.

Normalisierungsmethode. Sie können eine von fünf Optionen für die Normalisierung der Objektwerte und Variablen auswählen. In einer Analyse kann nur eine Normalisierungsmethode gleichzeitig verwendet werden.

- **Variablenprinzipal.** Bei dieser Option wird die Assoziation zwischen Variablen optimiert. Die Komponentenladungen (Korrelationen mit Hauptkomponenten wie Dimensionen und Objektwerten) stellen die Koordinaten der Variablen im Objektraum dar. Dies ist nützlich, wenn Sie hauptsächlich an der Korrelation zwischen den Variablen interessiert sind.
- **Objektprinzipal.** Bei dieser Option werden die Distanzen zwischen den Objekten optimiert. Dies ist nützlich, wenn Sie hauptsächlich an den Unterschieden oder Übereinstimmungen zwischen den Objekten interessiert sind.
- **Symmetrisch.** Verwenden Sie diese Normalisierungsmethode, wenn Sie hauptsächlich an den Beziehungen zwischen den Objekten und den Variablen interessiert sind.
- **Unabhängig.** Verwenden Sie diese Normalisierungsmethode, wenn Sie die Distanzen zwischen den Objekten und die Korrelationen zwischen den Variablen getrennt untersuchen möchten.
- **Benutzerdefiniert.** Sie können einen beliebigen reellen Wert aus dem abgeschlossenen Intervall [–1, 1] eingeben. Der Wert 1 entspricht der Methode “Objektprinzipal”, der Wert 0 der Methode “Symmetrisch” und der Wert –1 der Methode “Variablenprinzipal”. Bei Werten größer als –1 und kleiner als 1 wird der Eigenwert sowohl über Objekte als auch Variablen gestreut. Mit dieser Methode können Sie Biplots und Triplots genau an Ihre Bedürfnisse anpassen.

Kriterien. Sie können die maximale Anzahl von Iterationen festlegen, die in der Prozedur bei den Berechnungen durchlaufen werden. Außerdem können Sie einen Wert für das Konvergenzkriterium auswählen. Die Prozedur wird abgebrochen, wenn die Differenz der Gesamtanpassung zwischen den letzten zwei Iterationen kleiner als der Konvergenzwert ist oder die Maximalzahl von Iterationen erreicht wurde.

Diagramme beschriften mit. Hiermit können Sie festlegen, ob in den Diagrammen Variablen- und Wertlabels oder Variablennamen und Werte verwendet werden sollen. Sie können auch eine Höchstlänge für die Beschriftungen festlegen.

Diagrammdimensionen. Hiermit legen Sie fest, wie die Statistiken in der Ausgabe angezeigt werden.

- **Alle Dimensionen in der Lösung anzeigen.** Alle Dimensionen in der Lösung werden in einer Streudiagramm-Matrix angezeigt.
- **Anzahl der Dimensionen beschränken.** Die angezeigten Dimensionen werden auf geplottete Paare beschränkt. Wenn Sie die Dimensionen beschränken, müssen Sie die niedrigste und die höchste Dimension auswählen, die jeweils geplottet werden soll. Die niedrigste Dimension muss im Bereich von 1 bis zur Anzahl der Dimensionen in der Lösung minus 1 liegen und wird gegen höhere Dimensionen geplottet. Der Wert für die höchste Dimension muss im Bereich von 2 bis zur Anzahl der Dimensionen in der Lösung liegen und gibt die höchste Dimension an, die beim Plotten der Dimensionspaare verwendet wird. Diese Spezifikation gilt für alle angeforderten multidimensionalen Plots.

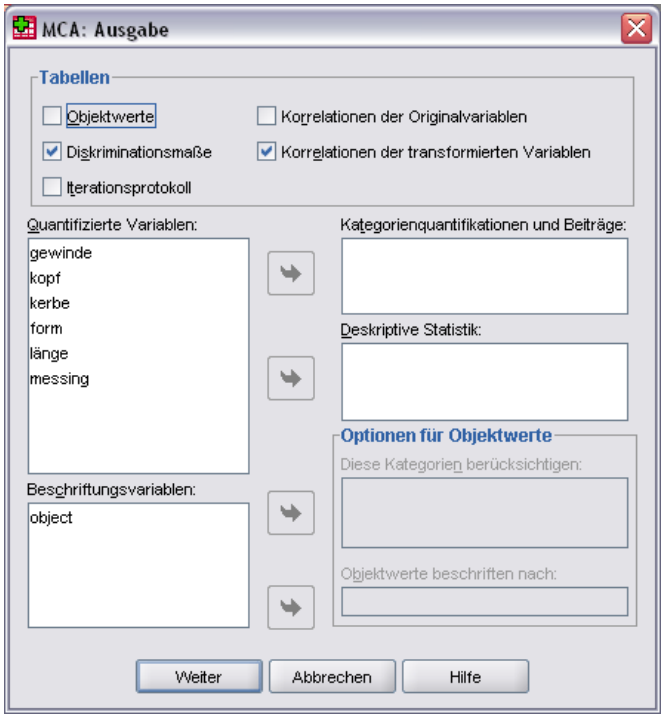
Konfiguration. Sie können die Koordinaten einer Konfiguration aus einer Datei einlesen. Die erste Variable in der Datei muss die Koordinaten der ersten Dimension, die zweite Variable die Koordinaten der zweiten Dimension usw. enthalten.

- **Initial.** Die in der Datei angegebene Konfiguration wird als Ausgangspunkt für die Analyse verwendet.
- **Fest.** Die in der Datei angegebene Konfiguration wird zum Anpassen der Variablen verwendet. Die Variablen, die angepaßt werden, müssen als Analysevariablen ausgewählt werden. Da die Konfiguration aber fest ist, werden Sie als Zusatzvariablen verarbeitet (und müssen deshalb nicht als Zusatzvariablen ausgewählt werden).

Mehrfachkorrespondenzanalyse: Ausgabe

Im Dialogfeld "Ausgabe" können Sie festlegen, dass Tabellen für Objektwerte, Diskriminationsmaße, das Iterationsprotokoll, Korrelationen der ursprünglichen und transformierten Variablen, Kategorienquantifikationen für ausgewählte Variablen und deskriptive Statistiken für ausgewählte Variablen erstellt werden sollen.

Abbildung 6-7
Dialogfeld "Ausgabe"



Objektwerte. Hiermit werden die Objektwerte (z. B. Masse, Trägheit, Beiträge) angezeigt. Die folgenden Optionen sind verfügbar:

- **Diese Kategorien berücksichtigen.** Hiermit werden die Kategorienummern der ausgewählten Analysevariablen angezeigt.
- **Objektwerte beschriften nach.** Sie können eine der in der Liste Beschriftungsvariablen enthaltenen Variablen auswählen, um die Objekte zu beschriften.

Diskriminationsmaße. Zeigt die Diskriminationsmaße pro Variable und Dimension an.

Iterationsprotokoll. Für jede Iteration werden berücksichtigte Varianz, Verlust sowie Steigerung der berücksichtigten Varianz ausgegeben.

Korrelationen der Originalvariablen. Hiermit werden die Korrelationsmatrix der ursprünglichen Variablen und die Eigenwerte dieser Matrix ausgegeben.

Korrelationen der transformierten Variablen. Hiermit werden die Korrelationsmatrix der transformierten (optimal skalierten) Variablen und die Eigenwerte dieser Matrix ausgegeben.

Kategorienquantifikationen und Beiträge. Dies sind die Kategoriequantifikationen und (Koordinaten) – darunter Masse, Trägheit und Beiträge – pro Dimension jeder ausgewählten Variable.

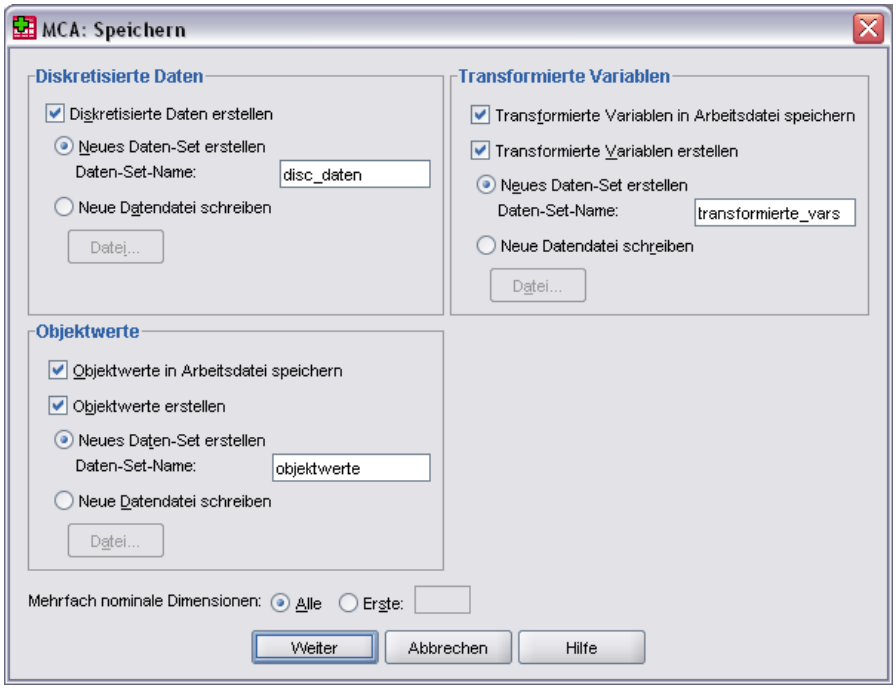
Deskriptive Statistik. Hiermit werden Häufigkeiten, Anzahl der fehlenden Werte und Modalwert für die ausgewählten Variablen ausgegeben.

Mehrfachkorrespondenzanalyse: Speichern

Im Dialogfeld “Speichern” können Sie festlegen, dass diskretisierte Daten, Objektwerte und transformierte Werte in einer SPSS-Datendatei oder in einem Daten-Set gespeichert werden. Außerdem können die transformierten Werte und die Objektwerte in der Arbeitsdatei gespeichert werden.

- Daten-Sets sind in der aktuellen Sitzung verfügbar. In nachfolgenden Sitzungen stehen Daten-Sets nur zur Verfügung, wenn Sie sie ausdrücklich als Datendateien speichern. Die Namen von Daten-Sets müssen den Regeln zum Benennen von Variablen entsprechen.
- Die Namen für die Dateien und Daten-Sets müssen sich bei allen Arten von gespeicherten Daten unterscheiden.
- Wenn Sie die Objektwerte oder die transformierten Werte in der Arbeitsdatei speichern, können Sie die Anzahl der mehrfach nominalen Dimensionen angeben.

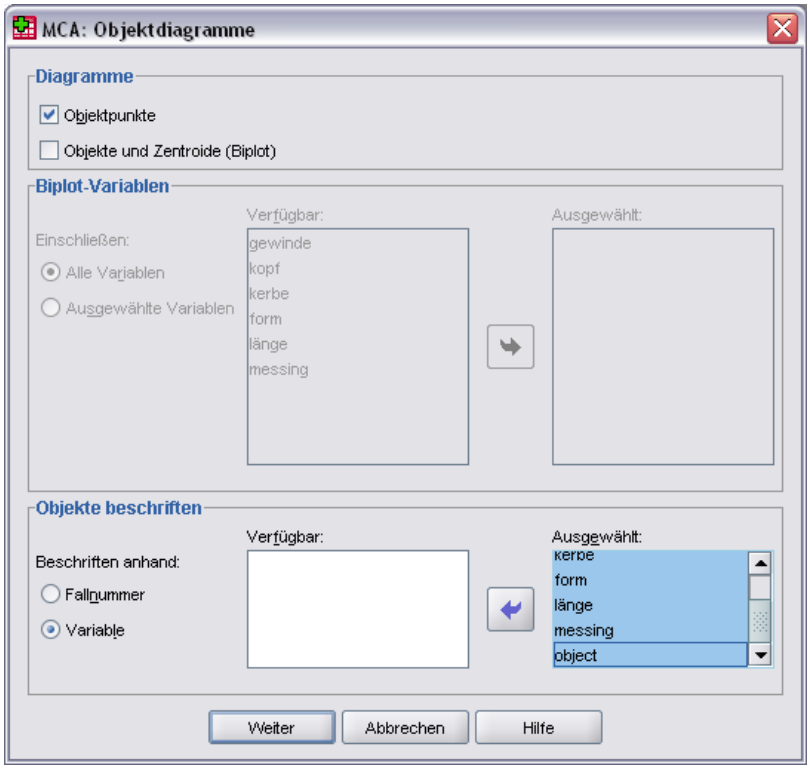
Abbildung 6-8
Dialogfeld "Speichern"



Mehrfachkorrespondenzanalyse: Objektdiagramme

Im Dialogfeld "Objektdiagramme" können Sie festlegen, welche Diagramme und für welche Variablen die Diagramme erstellt werden sollen.

Abbildung 6-9
Dialogfeld "Objektdiagramme"



Objektpunkte. Es wird ein Diagramm der Objektpunkte angezeigt.

Objekte und Zentroide (Biplot). Die Objektpunkte werden mit den Variablen-Zentroiden geplottet.

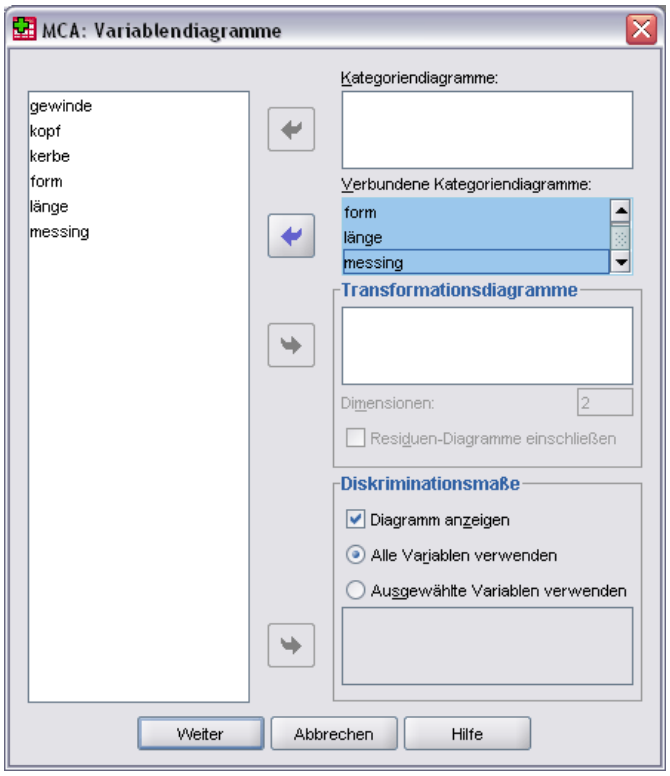
Biplot-Variablen. Die Biplots können auf allen Variablen oder einer Teilmenge der Variablen basieren.

Objekte beschriften. Objekte können mit den Kategorien ausgewählter Variablen (im Dialogfeld "Kategoriale Hauptkomponenten: Optionen" können Sie Kategorienummern oder Wertelabels wählen) oder den Fallnummern beschriftet werden. Wenn Variable ausgewählt ist, wird pro Variable ein Diagramm erstellt.

Mehrfachkorrespondenzanalyse: Variablendiagramme

Im Dialogfeld "Variablendiagramme" können Sie festlegen, welche Diagramme und für welche Variablen die Diagramme erstellt werden sollen.

Abbildung 6-10
Dialogfeld "Variablendiagramme"



Kategoriendiagramme. Für jede ausgewählte Variable wird ein Diagramm mit Zentroidkoordinaten erstellt. Die Kategorien werden in den Zentroiden der Objekte in den einzelnen Kategorien plaziert.

Verbundene Kategoriendiagramme. Hierbei handelt es sich um ein einzelnes Diagramm mit Zentroidkoordinaten jeder ausgewählten Variablen.

Transformationsdiagramme. Hierbei wird ein Diagramm der optimalen Kategorienquantifikationen gegen die Kategorienummern erstellt. Sie können die gewünschte Anzahl von Dimensionen angeben. Pro Dimension wird ein Diagramm erstellt. Sie können außerdem festlegen, dass für jede ausgewählte Variable ein Diagramm der Residuen ausgegeben wird.

Diskriminationsmaße. Erstellt ein einzelnes Diagramm der Diskriminationsmaße für die ausgewählten Variablen.

Zusätzliche Funktionen beim Befehl MULTIPLE CORRESPONDENCE

Sie können die Mehrfachkorrespondenzanalyse an Ihre Bedürfnisse anpassen, wenn Sie Ihre Auswahl in ein Syntax-Fenster einfügen und die resultierende Befehlssyntax für den Befehl `MULTIPLE CORRESPONDENCE` bearbeiten. Mit der Befehlssyntax-Sprache verfügen Sie außerdem über folgende Möglichkeiten:

- Mit dem Unterbefehl `SAVE` können Sie Stammnamen für die transformierten Variablen, Objektwerte und Approximationen festlegen, wenn Sie diese in der Arbeitsdatei speichern.

- Mit dem Unterbefehl PLOT kann die Höchstlänge für Beschriftungen für jedes Diagramm getrennt festgelegt werden.
- Mit dem Unterbefehl PLOT kann eine separate Variablenliste für Residuen-Diagramme festgelegt werden.

Vollständige Informationen zur Syntax finden Sie in der *Command Syntax Reference*.

Kapitel
7

Multidimensionale Skalierung (PROXSCAL)

Bei der multidimensionalen Skalierung wird versucht, die Struktur in einem Set von Ähnlichkeitsmaßen zwischen Objekten zu ermitteln. Dies wird durch das Zuweisen von Beobachtungen zu bestimmten Positionen in einem konzeptuellen Raum mit wenigen Dimensionen erzielt, wobei die Abstände zwischen den Punkten des Raumes mit den vorgegebenen Unähnlichkeiten bzw. Ähnlichkeiten so gut wie möglich übereinstimmen. Als Ergebnis werden die Objekte in diesem Raum mit wenigen Dimensionen mithilfe der Methode der kleinsten Quadrate dargestellt, was häufig zu einem besseren Verständnis der Daten beiträgt.

Beispiel. Die multidimensionale Skalierung kann das Ermitteln von Wahrnehmungsbeziehungen erheblich erleichtern. Der Hersteller eines Produkts kann zum Beispiel eine Umfrage durchführen, um das Produktimage zu ermitteln und ein Daten-Set zu erhalten, das die wahrgenommene Ähnlichkeit (Nähe) des Produkts mit Konkurrenzprodukten beschreibt. Unter Verwendung dieser Ähnlichkeitswerte und unabhängiger Variablen (z. B. des Preises) kann der Hersteller dann feststellen, welche Variablen für die Produktwahrnehmung durch die Verbraucher wichtig sind, und das Image des Produkts entsprechend ändern.

Statistiken und Diagramme. Iterationsprotokoll, Stressmaße, Stress-Zerlegung, Koordinaten des gemeinsamen Raums, Objektdistanzen innerhalb der endgültigen Konfiguration, individuelle Raumgewichtungen, individuelle Räume, transformierte Ähnlichkeiten, transformierte unabhängige Variablen, Stress-Diagramme, Streudiagramme des gemeinsamen Raums, Streudiagramme der individuellen Raumgewichtungen, Streudiagramme des individuellen Raums, Transformationsdiagramme, Shepard-Residuen-Diagramme und Transformationsdiagramme für unabhängige Variablen.

Daten. Daten können in Form von Distanz-Matrizen oder Variablen, die in Distanz-Matrizen umgewandelt werden, bereitgestellt werden. Die Matrizen können in einer Spalte oder über mehrere Spalten formatiert sein. Die Ähnlichkeiten können auf dem Skalierungsniveau der Verhältnisskala, Intervallskala, Ordinalskala oder Spline-Skala behandelt werden.

Annahmen. Es müssen mindestens drei Variablen angegeben werden. Die Anzahl der Dimensionen darf die Anzahl der Objekte minus 1 nicht übersteigen. Bei der Kombination mit mehreren Zufallsstarts wird die Dimensionsreduktion ausgelassen. Wenn nur eine Quelle angegeben ist, sind alle Modelle äquivalent zum Identitätsmodell. In diesem Fall wird für die Analyse standardmäßig das Identitätsmodell verwendet.

Verwandte Prozeduren. Eine Skalierung aller Variablen auf numerischem Niveau entspricht der standardmäßigen multidimensionalen Skalierung.

So berechnen Sie eine multidimensionale Skalierung:

▶ Wählen Sie die folgenden Befehle aus den Menüs aus:

Analysieren
  Skalieren
    Multidimensionale Skalierung (PROXSCAL)...

Daraufhin wird das Dialogfeld "Datenformat" geöffnet.

Abbildung 7-1
Dialogfeld "Datenformat"

▶ Geben Sie das Format der Daten an:

Datenformat. Geben Sie an, ob die Daten aus Ähnlichkeitsmaßen bestehen oder ob Sie aus den Daten Ähnlichkeitsmaße erstellen möchten.

Anzahl der Quellen. Wenn es sich bei den Daten um Ähnlichkeiten handelt, geben Sie an, ob eine Quelle oder mehrere Quellen für Ähnlichkeitsmaße vorhanden sind.

Eine Quelle. Wenn nur eine Quelle für Ähnlichkeiten vorhanden ist, geben Sie die Art der Formatierung des Daten-Sets an: Die Ähnlichkeiten befinden sich in einer Matrix über mehrere Spalten verteilt oder in einer einzigen Spalte mit zwei verschiedenen Variablen zur Kennzeichnung der Zeile und Spalte jeder Ähnlichkeit.

- **Ähnlichkeiten stammen aus mehreren Spalten einer Matrix.** Die Ähnlichkeits-Matrix erstreckt sich über mehrere Spalten, deren Anzahl der Anzahl der Objekte entspricht. Hiermit öffnen Sie das Dialogfeld "Ähnlichkeiten in Matrizen über mehrere Spalten".
- **Ähnlichkeiten stammen aus einer einzigen Spalte.** Die Ähnlichkeits-Matrix befindet sich in einer einzigen Spalte (bzw. Variablen). Es sind zwei weitere Variablen erforderlich, welche die Zeile und Spalte jeder Zelle angeben. Hiermit öffnen Sie das Dialogfeld "Ähnlichkeiten in einer einzigen Spalte".

Mehrere Quellen. Wenn mehrere Quellen für Ähnlichkeiten vorhanden sind, geben Sie die Art der Formatierung des Daten-Sets an: Die Ähnlichkeiten befinden sich über mehrere Spalten verteilt in gestapelten Matrizen, in mehreren Spalten mit einer Quelle pro Spalte oder in einer einzigen Spalte.

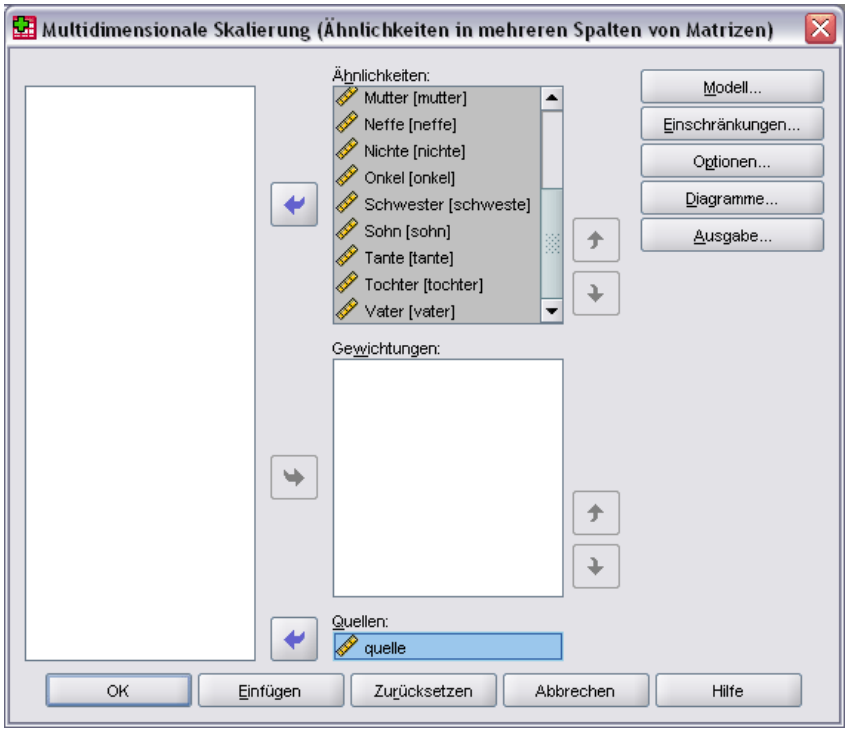
- **Ähnlichkeiten stammen aus mehreren Spalten in gestapelten Matrizen.** Die Ähnlichkeits-Matrizen erstrecken sich über mehrere Spalten, deren Anzahl der Anzahl der Objekte entspricht, und sind übereinandergestapelt in mehreren Zeilen, deren Anzahl der Anzahl der Objekte mal der Anzahl der Quellen entspricht. Hiermit öffnen Sie das Dialogfeld "Ähnlichkeiten in Matrizen über mehrere Spalten".
- **Ähnlichkeiten stammen aus Spalten (eine Quelle pro Spalte).** Die Ähnlichkeits-Matrizen sind in mehreren Spalten (bzw. Variablen) zusammengefasst. Es sind zwei weitere Variablen erforderlich, welche die Zeile und Spalte jeder Zelle angeben. Hiermit öffnen Sie das Dialogfeld "Ähnlichkeiten in Spalten".
- **Ähnlichkeiten sind in einer Spalte gestapelt.** Die Ähnlichkeits-Matrizen sind in einer einzigen Spalte (bzw. Variablen) zusammengefasst. Es sind drei weitere Variablen erforderlich, welche die Zeile, Spalte und Quelle jeder Zelle angeben. Hiermit öffnen Sie das Dialogfeld "Ähnlichkeiten in einer einzigen Spalte".

▶ Klicken Sie auf Definieren.

Ähnlichkeiten in mehreren Spalten von Matrizen

Wenn Sie im Dialogfeld “Datenformat” das Datenmodell mit Ähnlichkeiten in Matrizen auswählen (aus einer oder mehreren Quellen), wird das Hauptdialogfeld folgendermaßen dargestellt:

Abbildung 7-2
Dialogfeld “Ähnlichkeiten in mehreren Spalten von Matrizen”



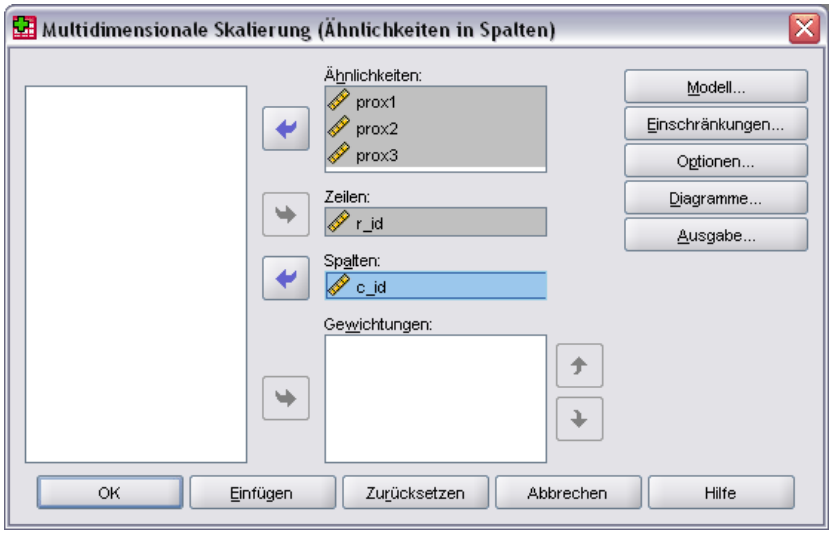
- Wählen Sie mindestens drei Ähnlichkeitsvariablen aus. (Achten Sie darauf, dass die Reihenfolge der Variablen in der Liste der Reihenfolge der Spalten mit den Ähnlichkeiten entspricht.)
- Wählen Sie optional Gewichtungsvariablen aus. Die Anzahl muss der Anzahl der ausgewählten Ähnlichkeitsvariablen entsprechen. (Achten Sie darauf, dass die Reihenfolge der Gewichtungen der Reihenfolge der damit gewichteten Ähnlichkeiten entspricht.)
- Wenn mehrere Quellen vorhanden sind, wählen Sie optional eine Quellenvariable aus. (Die Anzahl der Fälle in jeder Ähnlichkeitsvariablen muss der Anzahl der Ähnlichkeitsvariablen multipliziert mit der Anzahl der Quellen entsprechen.)

  Darüber hinaus können Sie ein Modell für die multidimensionale Skalierung angeben, Einschränkungen für den gemeinsamen Raum festlegen, Konvergenzkriterien festlegen, die zu verwendende Ausgangskonfiguration angeben sowie Diagramme und die Ausgabe auswählen.

Ähnlichkeiten in Spalten

Wenn Sie im Dialogfeld “Datenformat” das Modell mit mehreren Spalten für mehrere Quellen auswählen, wird das Hauptdialogfeld folgendermaßen dargestellt:

Abbildung 7-3
Dialogfeld “Ähnlichkeiten in Spalten”



- Wählen Sie mindestens zwei Ähnlichkeitsvariablen aus. (Es wird davon ausgegangen, dass es sich bei den Variablen um Distanzmatrizen aus verschiedenen Quellen handelt.)
- Wählen Sie eine Zeilenvariable zum Festlegen der Zeilenpositionen für die Ähnlichkeiten in jeder Ähnlichkeitsvariablen aus.
- Wählen Sie eine Spaltenvariable zum Festlegen der Spaltenpositionen für die Ähnlichkeiten in jeder Ähnlichkeitsvariablen aus. (Zellen der Distanzmatrix, denen keine Zeilen-/Spaltenposition zugewiesen ist, werden als fehlend behandelt.)
- Wählen Sie optional Gewichtungsvariablen aus. Die Anzahl muss der Anzahl der ausgewählten Ähnlichkeitsvariablen entsprechen.

  Darüber hinaus können Sie ein Modell für die multidimensionale Skalierung angeben, Einschränkungen für den gemeinsamen Raum festlegen, Konvergenzkriterien festlegen, die zu verwendende Ausgangskonfiguration angeben sowie Diagramme und die Ausgabe auswählen.

Ähnlichkeiten in einer Spalte

Wenn Sie im Dialogfeld “Datenformat” das Ein-Spalten-Modell auswählen (entweder aus einer oder mehreren Quellen), wird das Hauptdialogfeld folgendermaßen dargestellt:

Abbildung 7-4
Dialogfeld “Ähnlichkeiten in einer Spalte”

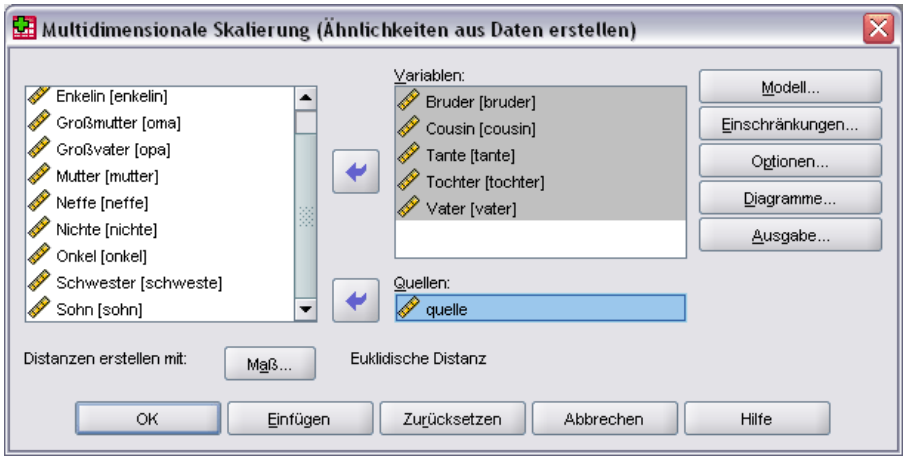
- Wählen Sie eine Ähnlichkeitsvariable aus. (Es wird davon ausgegangen, dass es sich dabei um eine oder mehrere Distanzmatrizen handelt.)
- Wählen Sie eine Zeilenvariable zum Festlegen der Zeilenpositionen für die Ähnlichkeiten in der Ähnlichkeitsvariablen aus.
- Wählen Sie eine Spaltenvariable zum Festlegen der Spaltenpositionen für die Ähnlichkeiten in der Ähnlichkeitsvariablen aus.
- Wenn mehrere Quellen vorhanden sind, wählen Sie eine Quellenvariable aus. (Zellen der Distanzmatrix, denen keine Zeilen-/Spaltenposition zugewiesen ist, werden für jede Quelle als fehlend behandelt.)
- Wählen Sie optional eine Gewichtungsvariable aus.

  Darüber hinaus können Sie ein Modell für die multidimensionale Skalierung angeben, Einschränkungen für den gemeinsamen Raum festlegen, Konvergenzkriterien festlegen, die zu verwendende Ausgangskonfiguration angeben sowie Diagramme und die Ausgabe auswählen.

Ähnlichkeiten aus Daten erstellen

Wenn Sie im Dialogfeld “Datenformat” die Option zum Erstellen von Ähnlichkeiten aus Daten auswählen, wird das Hauptdialogfeld folgendermaßen dargestellt:

Abbildung 7-5
Dialogfeld “Ähnlichkeiten aus Daten erstellen”

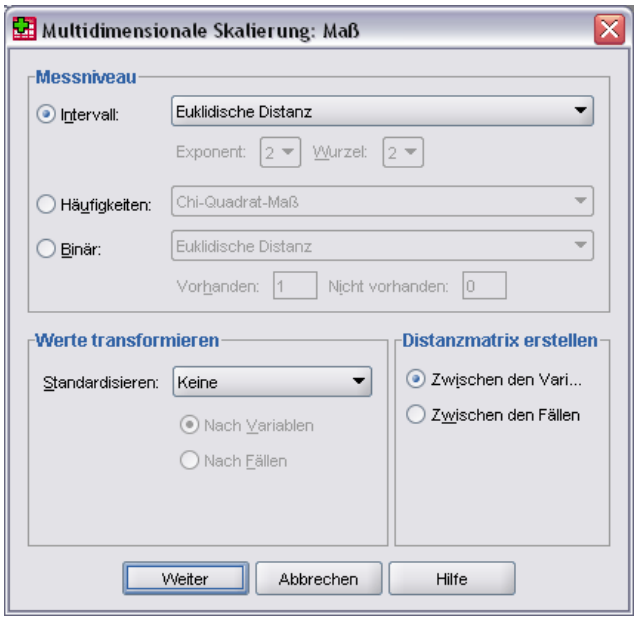


- Wenn Sie Distanzen zwischen Variablen erzeugen (siehe Dialogfeld “Distanzen aus Daten erstellen”), wählen Sie mindestens drei Variablen aus. Diese Variablen dienen zum Erstellen der Distanzmatrix (oder Distanzmatrizen, wenn mehrere Quellen vorhanden sind). Bei der Erzeugung von Distanzen zwischen Fällen wird nur eine Variable benötigt.
- Wenn mehrere Quellen vorhanden sind, wählen Sie eine Quellenvariable aus.
- Wählen Sie bei Bedarf ein Maß für die Erstellung der Ähnlichkeiten aus.

  Darüber hinaus können Sie ein Modell für die multidimensionale Skalierung angeben, Einschränkungen für den gemeinsamen Raum festlegen, Konvergenzkriterien festlegen, die zu verwendende Ausgangskonfiguration angeben sowie Diagramme und die Ausgabe auswählen.

Distanzen aus Daten erstellen

Abbildung 7-6
Dialogfeld "Distanzen aus Daten erstellen"



Die multidimensionale Skalierung verwendet Unähnlichkeitsdaten, um eine Skalierungslösung zu erstellen. Wenn Ihre Daten multivariate Daten darstellen (Werte gemessener Variablen), müssen Sie Unähnlichkeitsdaten erstellen, um eine multidimensionale Skalierungslösung berechnen zu können. Sie können Optionen für das Erstellen von Unähnlichkeitsmaßen aus Ihren Daten festlegen.

Messniveau. Hier können Sie das Unähnlichkeitsmaß für Ihre Analyse festlegen. Wählen Sie aus dem Gruppenfeld "Maß" die Option aus, die Ihrem Datentyp entspricht. Wählen Sie dann aus dem Dropdown-Listenfeld ein Maß aus, das diesem Messwerttyp entspricht. Die folgenden Optionen sind verfügbar:

- **Intervall**. Euklidischer Abstand, quadrierter Euklidischer Abstand, Tschebyscheff, Block, Minkowski oder ein benutzerdefiniertes Maß.
- **Häufigkeiten**. Chi-Quadrat-Maß oder Phi-Quadrat-Maß.
- **Binär**. Euklidischer Abstand, quadrierter Euklidischer Abstand, Größendifferenz, Musterdifferenz, Varianz und Distanzmaß nach Lance und Williams.

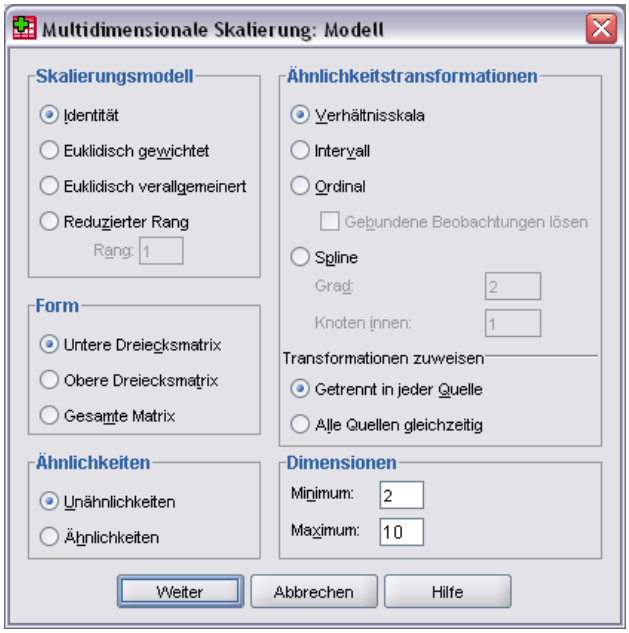
Distanzmatrix erstellen. Mit dieser Funktion können Sie die Einheit der Analyse wählen. Zur Auswahl stehen "Zwischen den Variablen" oder "Zwischen den Fällen".

Werte transformieren. In bestimmten Fällen, zum Beispiel wenn die Variablen mit sehr unterschiedlichen Skalen gemessen werden, empfiehlt sich das Standardisieren der Werte vor dem Berechnen der Ähnlichkeiten (nicht auf binäre Daten anwendbar). Wählen Sie eine Standardisierungsmethode aus dem Dropdown-Listenfeld "Standardisieren" aus. (Wenn keine Standardisierung erforderlich ist, wählen Sie Keine aus).

Definieren eines Modells für die multidimensionale Skalierung

Im Dialogfeld "Modell" können Sie ein Skalierungsmodell, die minimale und maximale Anzahl seiner Dimensionen, die Struktur der Distanzmatrix und die für die Ähnlichkeiten zu verwendende Transformation angeben sowie festlegen, ob Ähnlichkeiten für jede Quelle getrennt oder unabhängig von der Quelle transformiert werden.

Abbildung 7-7
Dialogfeld "Modell"



Skalierungsmodell. Wählen Sie eine der folgenden Alternativen aus:

- **Identität.** Alle Quellen weisen dieselbe Konfiguration auf.
- **Euklidisch gewichtet.** Hierbei handelt es sich um ein Modell für individuelle Differenzen. Jede Quelle hat einen individuellen Raum, in dem jede Dimension des gemeinsamen Raums unterschiedlich gewichtet ist.
- **Euklidisch verallgemeinert.** Hierbei handelt es sich um ein Modell für individuelle Differenzen. Jede Quelle hat einen individuellen Raum, der einer Rotation des gemeinsamen Raums und einer darauf folgenden differenzierten Gewichtung der Dimensionen entspricht.
- **Reduzierter Rang.** Bei diesem Modell handelt es sich um ein verallgemeinertes euklidisches Modell, für das Sie den Rang des individuellen Raums angeben können. Der angegebene Rang muss größer oder gleich 1 und kleiner als die Höchstanzahl von Dimensionen sein.

Form. Geben Sie an, ob die Ähnlichkeiten der unteren oder oberen Dreiecksmatrix der Distanzmatrix entnommen werden sollen. Sie können angeben, dass die gesamte Matrix verwendet werden soll. In diesem Fall wird die gewichtete Summe der unteren und oberen Dreiecksmatrix analysiert. Obwohl nur die angegebenen Abschnitte verwendet werden, muss in jeden Fall die gesamte Matrix einschließlich der Diagonalen angegeben werden.

Ähnlichkeiten. Geben Sie an, ob die Distanzmatrix Ähnlichkeits- oder Unähnlichkeitsmaße enthält.

Ähnlichkeitstransformationen. Wählen Sie eine der folgenden Alternativen aus:

- **Verhältnisskala.** Die transformierten Ähnlichkeiten sind proportional zu den ursprünglichen Ähnlichkeiten. Diese Option ist nur für Ähnlichkeiten mit positiven Werten zulässig.
- **Intervall.** Die transformierten Ähnlichkeiten sind proportional zu den ursprünglichen Ähnlichkeiten plus einem konstanten Term. Durch die Konstante wird sichergestellt, dass alle Ähnlichkeiten positive Werte haben.
- **Ordinal.** Die transformierten Ähnlichkeiten weisen dieselbe Ordnung wie die ursprünglichen Ähnlichkeiten auf. Sie geben an, ob die Bindung von gebundenen Ähnlichkeiten erhalten bleiben muss oder aufgehoben werden darf.
- **Spline.** Die transformierten Ähnlichkeiten sind eine geglättete, nicht fallende, stückweise, polynomiale Transformation der ursprünglichen Ähnlichkeiten. Sie geben den Grad des Polynoms und die Anzahl der inneren Knoten an.

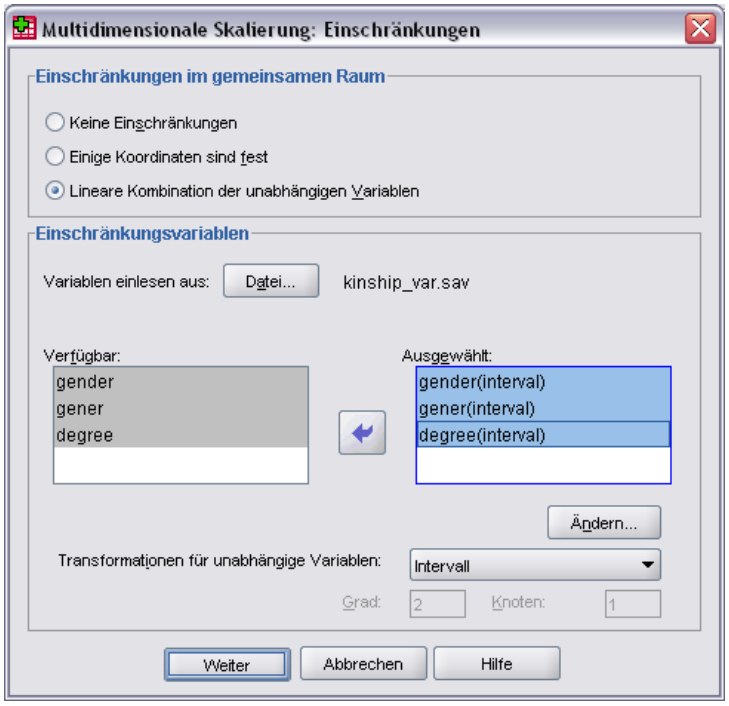
Transformationen zuweisen. Geben Sie an, ob nur Ähnlichkeiten innerhalb jeder Quelle miteinander verglichen werden oder ob die Vergleiche unabhängig von der Quelle sein sollen.

Dimensionen. Standardmäßig wird eine Lösung in zwei Dimensionen berechnet (Minimum = 2, Maximum = 2). Geben Sie für das Minimum und Maximum eine beliebige ganze Zahl zwischen 1 und der Anzahl der Objekte minus 1 an (solange das Minimum kleiner oder gleich dem Maximum ist). Die Prozedur berechnet eine Lösung in der höchsten Dimension und reduziert die Dimensionalität dann schrittweise, bis die niedrigste Dimensionalität erreicht ist.

Multidimensionale Skalierung: Einschränkungen

Im Dialogfeld “Einschränkungen” können Sie Einschränkungen für den gemeinsamen Raum festlegen.

Abbildung 7-8
Dialogfeld "Einschränkungen"



Einschränkungen im gemeinsamen Raum. Geben Sie die gewünschte Art der Einschränkung an.

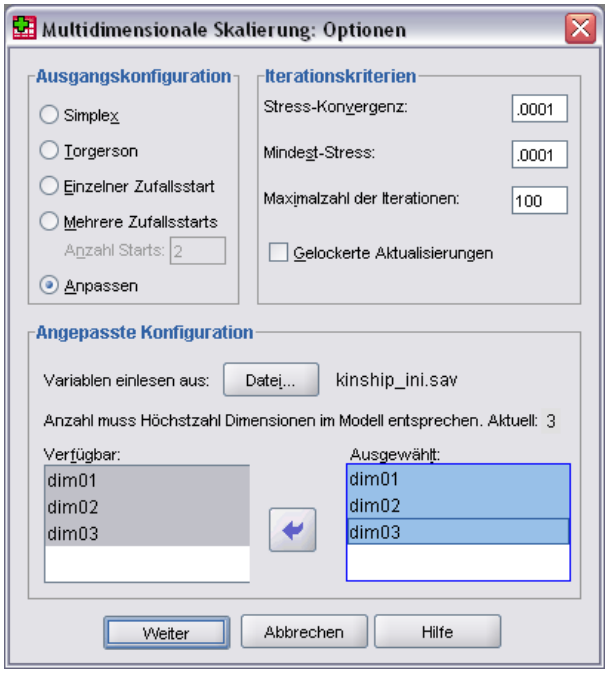
- **Keine Einschränkungen.** Es werden keine Einschränkungen für den gemeinsamen Raum festgelegt.
- **Einige Koordinaten sind fest.** Die erste ausgewählte Variable enthält die Koordinaten der Objekte in der ersten Dimension, die zweite Variable entspricht den Koordinaten der zweiten Dimension usw. Ein fehlender Wert zeigt an, dass eine Koordinate in einer Dimension frei ist. Die Anzahl der ausgewählten Variablen muss der angegebenen Höchstanzahl von Dimensionen entsprechen.
- **Lineare Kombination der unabhängigen Variablen.** Der gemeinsame Raum muss eine lineare Kombination der ausgewählten Variablen sein.

Einschränkungsvariablen. Wählen Sie die Variablen aus, die die Einschränkungen für den gemeinsamen Raum definieren. Wenn Sie eine lineare Kombination angegeben haben, geben Sie für die Einschränkungsvariablen eine Intervall-, Nominal-, Ordinal- oder Spline-Transformation an. Die Anzahl der Fälle muss in jedem Fall der Anzahl der Objekte entsprechen.

Multidimensionale Skalierung: Optionen

Das Dialogfeld "Optionen" ermöglicht Ihnen die Auswahl der Anfangskonfiguration, die Angabe von Iterations- und Konvergenzkriterien und die Wahl zwischen standardmäßigen und gelockerten Aktualisierungen.

Abbildung 7-9
Dialogfeld "Optionen"



Ausgangskonfiguration. Wählen Sie eine der folgenden Optionen:

- **Simplex.** Objekte werden in der höchsten Dimension in gleichen Abständen voneinander plaziert. Eine Iteration wird durchgeführt, um diese Konfiguration der höchsten Dimension zu optimieren, und anschließend wird eine Dimensionsreduktion vorgenommen, um eine Ausgangskonfiguration mit der maximalen, im Dialogfeld "Modell" angegebenen Anzahl von Dimensionen zu erhalten.
- **Torgerson.** Als Ausgangskonfiguration wird eine klassische Skalierungslösung verwendet.
- **Einzelner Zufallsstart.** Eine Konfiguration wird zufällig ausgewählt.
- **Mehrere Zufallsstarts.** Es werden mehrere Konfigurationen nach dem Zufallsprinzip ausgewählt. Die Konfiguration mit dem niedrigsten normalisierten Roh-Stress wird als Ausgangskonfiguration verwendet.
- **Anpassen.** Wählen Sie Variablen aus, die die Koordinaten Ihrer eigenen Ausgangskonfiguration enthalten. Die Anzahl der ausgewählten Variablen muss der angegebenen Höchstanzahl von Dimensionen entsprechen, wobei die erste Variable den Koordinaten von Dimension 1, die zweite Variable den Koordinaten von Dimension 2 entspricht usw. Die Anzahl der Fälle in jeder Variablen muss der Anzahl der Objekte entsprechen.

Iterationskriterien. Geben Sie Werte für Iterationskriterien an.

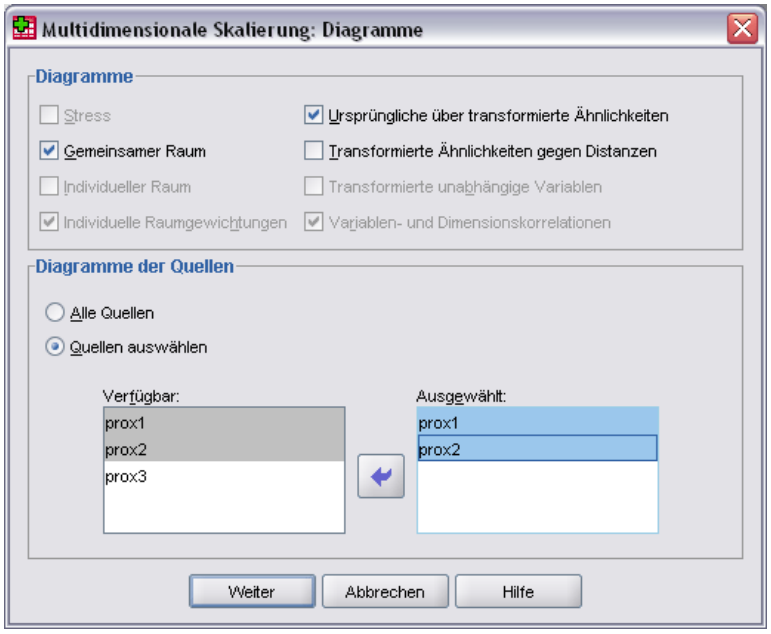
- **Stress-Konvergenz.** Der Algorithmus führt keine weiteren Iterationen durch, wenn die Differenz von aufeinander folgenden normalisierten Roh-Stress-Werten unter dem hier angegebenen Wert liegt, der zwischen 0,0 und 1,0 liegen muss.
- **Mindest-Stress.** Der Algorithmus wird beendet, wenn der normalisierte Roh-Stress unter den hier angegebenen Wert fällt, der zwischen 0,0 und 1,0 liegen muss.

- **Maximalzahl der Iterationen.** Der Algorithmus führt die angegebene Anzahl von Iterationen durch, falls nicht vorher eines der oben genannten Kriterien erfüllt wird.
- **Gelockerte Aktualisierungen.** Gelockerte Aktualisierungen beschleunigen den Algorithmus. Diese Aktualisierungen können mit keinem anderen Modell als dem Identitätsmodell und nicht zusammen mit Einschränkungen verwendet werden.

Multidimensionale Skalierung: Diagramme, Version 1

Im Dialogfeld "Diagramme" können Sie die zu erstellenden Diagramme festlegen. Wenn das Datenformat "Ähnlichkeiten in Spalten" ausgewählt wurde, wird das Dialogfeld "Diagramme" wie folgt angezeigt. Geben Sie für die Diagramme Individuelle Raumgewichtungen, Ursprüngliche über transformierte Ähnlichkeiten und Transformierte Ähnlichkeiten über Distanzen die Quellen an, für die die Diagramme erzeugt werden sollen. Die Liste der verfügbaren Quellen entspricht der Liste der Ähnlichkeitsvariablen im Hauptdialogfeld.

Abbildung 7-10
Dialogfeld "Diagramme", Version 1



Stress. Ein Diagramm wird erstellt, in dem der normalisierte Roh-Stress über die Dimensionen abgebildet wird. Dieses Diagramm wird nur erstellt, wenn die Höchstanzahl von Dimensionen größer als die Mindestanzahl von Dimensionen ist.

Gemeinsamer Raum. Eine Streudiagramm-Matrix der Koordinaten des gemeinsamen Raums wird angezeigt.

Individueller Raum. Für jede Quelle werden die Koordinaten der individuellen Räume in Streudiagramm-Matrizen angezeigt. Dies ist nur möglich, wenn im Dialogfeld "Modell" eines der Modelle für individuelle Differenzen angegeben ist.

Individuelle Raumgewichtungen. Ein Streudiagramm der individuellen Raumgewichtungen wird erstellt. Dies ist nur möglich, wenn im Dialogfeld “Modell” eines der Modelle für individuelle Differenzen angegeben ist. Für das Modell “Euklidisch gewichtet” werden die Gewichtungen in Diagrammen mit einer Dimension pro Achse ausgegeben. Für das Modell “Euklidisch verallgemeinert” wird ein Diagramm pro Dimension erstellt, das Rotation und Gewichtung der jeweiligen Dimension anzeigt. Das für das Modell “Reduzierter Rang” erstellte Diagramm gleicht dem Diagramm für das Modell “Euklidisch verallgemeinert”, die Anzahl der Dimensionen für die individuellen Räume ist jedoch reduziert.

Ursprüngliche über transformierte Ähnlichkeiten. Es werden Diagramme der ursprünglichen Ähnlichkeiten im Vergleich zu den transformierten Ähnlichkeiten erstellt.

Transformierte Ähnlichkeiten über Distanzen. Die transformierten Ähnlichkeiten werden über die Distanzen abgebildet.

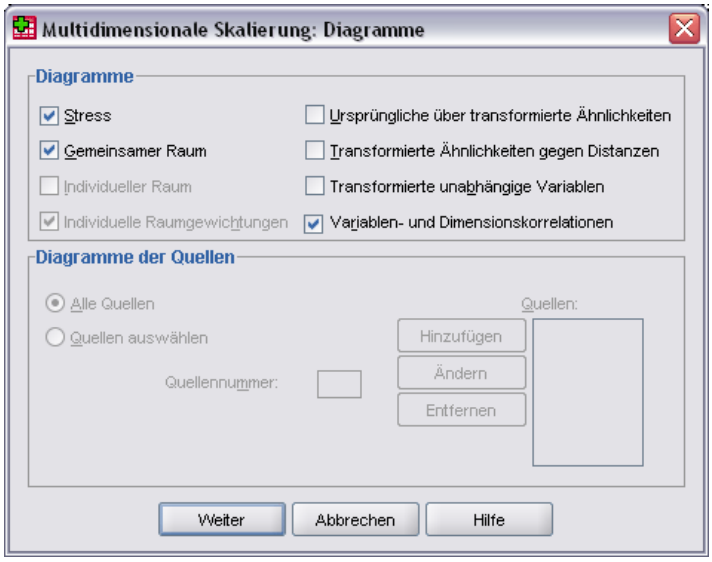
Transformierte unabhängige Variablen. Es werden Transformationsdiagramme für die unabhängigen Variablen erstellt.

Variablen- und Dimensionskorrelationen. Es wird ein Diagramm der Korrelationen zwischen den unabhängigen Variablen und den Dimensionen des gemeinsamen Raums angezeigt.

Multidimensionale Skalierung: Diagramme, Version 2

Im Dialogfeld “Diagramme” können Sie die zu erstellenden Diagramme festlegen. Wenn ein beliebiges Datenformat außer dem Datenformat “Ähnlichkeiten in Spalten” ausgewählt wurde, wird das Dialogfeld “Diagramme” wie folgt angezeigt. Geben Sie für die Diagramme Individuelle Raumgewichtungen, Ursprüngliche über transformierte Ähnlichkeiten und Transformierte Ähnlichkeiten über Distanzen die Quellen an, für die die Diagramme erzeugt werden sollen. Die eingegebenen Quellennummern müssen Werte der im Hauptdialogfeld angegebenen Quellenvariable sein und im Bereich von 1 bis zur Anzahl der Quellen liegen.

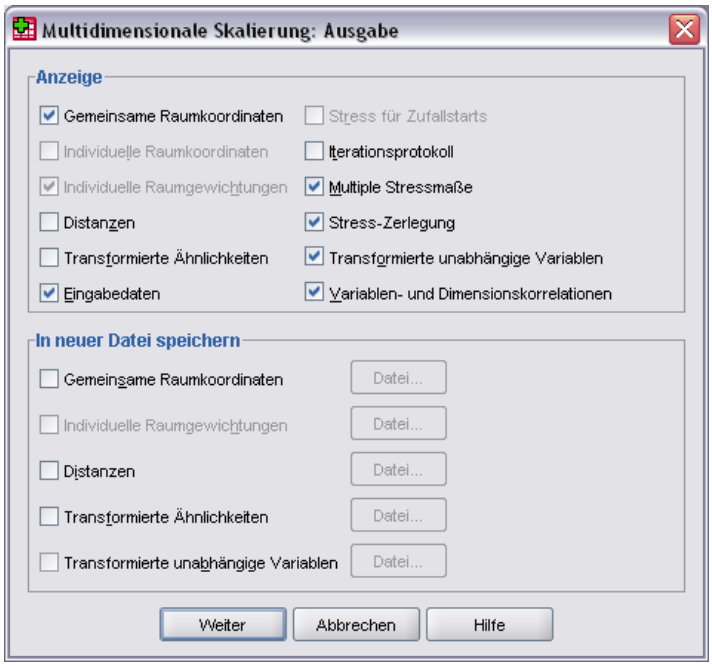
Abbildung 7-11
Dialogfeld “Diagramme”, Version 2



Multidimensionale Skalierung: Ausgabe

Im Dialogfeld “Ausgabe” können Sie die angezeigte Ausgabe steuern und Auszüge davon in eigenen Dateien speichern.

Abbildung 7-12
Dialogfeld “Ausgabe”



Anzeige. Wählen Sie mindestens eine der folgenden Anzeigeoptionen aus:

- **Gemeinsame Raumkoordinaten.** Zeigt die Koordinaten des gemeinsamen Raums an.
- **Individuelle Raumkoordinaten.** Zeigt nur dann die Koordinaten der individuellen Räume an, wenn das Modell kein Identitätsmodell ist.
- **Individuelle Raumgewichtungen.** Zeigt nur dann die individuellen Raumgewichtungen an, wenn eines der Modelle für individuelle Differenzen angegeben ist. Je nach Modell werden die Raumgewichtungen in Rotations- und Dimensionsgewichtungen zerlegt, die ebenfalls angezeigt werden.
- **Distanzen.** Zeigt die Distanzen zwischen den Objekten in der Konfiguration an.
- **Transformierte Ähnlichkeiten.** Zeigt die transformierten Ähnlichkeiten zwischen den Objekten in der Konfiguration an.
- **Eingabedaten.** Schließt die ursprünglichen Ähnlichkeiten und, sofern vorhanden, die Datengewichtungen, die Ausgangskonfiguration und die festen Koordinaten der unabhängigen Variablen ein.
- **Stress für Zufallsstarts.** Zeigt den Startwert für Zufallszahlen und den normalisierten Roh-Stress-Wert für jeden Zufallsstart an.
- **Iterationsprotokoll.** Zeigt das Iterationsprotokoll des Hauptalgorithmus an.

- **Multiple Stressmaße.** Zeigt verschiedene Stress-Werte an. Die Tabelle enthält Werte für normalisierten Roh-Stress, Stress-I, Stress-II, S-Stress, erklärte Streuung und den Kongruenzkoeffizienten nach Tucker.
- **Stress-Zerlegung.** Zeigt eine Objekte- und Quellenzerlegung des endgültigen normalisierten Roh-Stress einschließlich des Durchschnitts pro Objekt und des Durchschnitts pro Quelle an.
- **Transformierte unabhängige Variablen.** Wenn eine Einschränkung der linearen Kombination ausgewählt wurde, werden die transformierten unabhängigen Variablen und die entsprechenden Regressionsgewichtungen angezeigt.
- **Variablen- und Dimensionskorrelationen.** Wenn eine Einschränkung der linearen Kombination ausgewählt wurde, werden die Korrelationen zwischen den unabhängigen Variablen und den Dimensionen des gemeinsamen Raums angezeigt.

In neuer Datei speichern. Sie können die Koordinaten des gemeinsamen Raums, die individuellen Raumgewichtungen, die Distanzen, die transformierten Ähnlichkeiten und die transformierten unabhängigen Variablen in eigenen SPSS-Datendateien speichern.

Zusätzliche Funktionen beim Befehl PROXSCAL

Sie können die kategoriale Hauptkomponentenanalyse individuell anpassen, indem Sie die Auswahl in ein Syntax-Fenster einfügen und die daraus resultierende Befehlssyntax für `PROXSCAL` bearbeiten. Mit der Befehlssyntax-Sprache verfügen Sie außerdem über folgende Möglichkeiten:

- Geben Sie eigene Variablenlisten für Transformationen und Residuen-Diagramme an (mit dem Unterbefehl `PLOT`).
- Geben Sie eigene Quelllisten für individuelle Raumgewichtungen, Transformationen und Residuen-Diagramme an (mit dem Unterbefehl `PLOT`).
- Geben Sie eine Untergruppe der Transformationsdiagramme für unabhängige Variablen an, die angezeigt werden sollen (mit dem Unterbefehl `PLOT`).

Vollständige Informationen zur Syntax finden Sie in der *Command Syntax Reference*.

Kapitel
8

Multidimensionale Entfaltung (PREFSCAL)

In der Prozedur “Multidimensionale Entfaltung” wird versucht, eine gemeinsame, quantitative Skala zu finden, mit der Sie die Beziehung zwischen zwei Gruppen von Objekten visuell untersuchen können.

Beispiele. Sie haben 21 Personen geben, 15 Frühstücksartikel in der Vorzugsreihenfolge von 1 bis 15 zu ordnen. Mit der Prozedur “Multidimensionale Entfaltung” können Sie ermitteln, dass die Personen die Frühstücksartikel in erster Linie auf zwei Arten unterscheiden: zwischen Weiß- und Vollkornbrot und zwischen kalorienhaltigen und -armen Artikeln.

In einer anderen Umfrage haben Sie eine Gruppe von Fahrern gebeten, 26 Fahrzeugmodelle anhand von 10 Attributen auf einer 6-Punkte-Skala von “1=trifft überhaupt nicht zu” bis “6=trifft vollständig zu” zu bewerten. Die Werte werden über die einzelnen Personen gemittelt und als Ähnlichkeiten verwendet. Mit der Prozedur “Multidimensionale Entfaltung” finden Sie Gruppen von ähnlichen Modellen sowie von Attributen, mit denen diese am ehesten in Verbindung gebracht werden.

Statistiken und Diagramme. Mit der Prozedur “Multidimensionale Entfaltung” können Sie ein Iterationsprotokoll, Stressmaße, Stress-Zerlegung, Koordinaten des gemeinsamen Raums, Objektdistanzen innerhalb der endgültigen Konfiguration, individuelle Raumgewichtungen, individuelle Räume, transformierte Distanzen, Stress-Diagramme, Streudiagramme des gemeinsamen Raums, Streudiagramme der individuellen Raumgewichtungen, Streudiagramme des individuellen Raums, Transformationsdiagramme und Shepard-Residuen-Diagramme erstellen.

Daten. Die Daten werden in Form von rechtwinkligen Distanzmatrizen bereitgestellt. Jede Spalte wird als separates Spaltenobjekt betrachtet. Jede Zeile einer Distanzmatrix wird als separates Zeilenobjekt betrachtet. Wenn es mehrere Quellen für die Distanzen gibt, werden die Matrizen gestapelt.

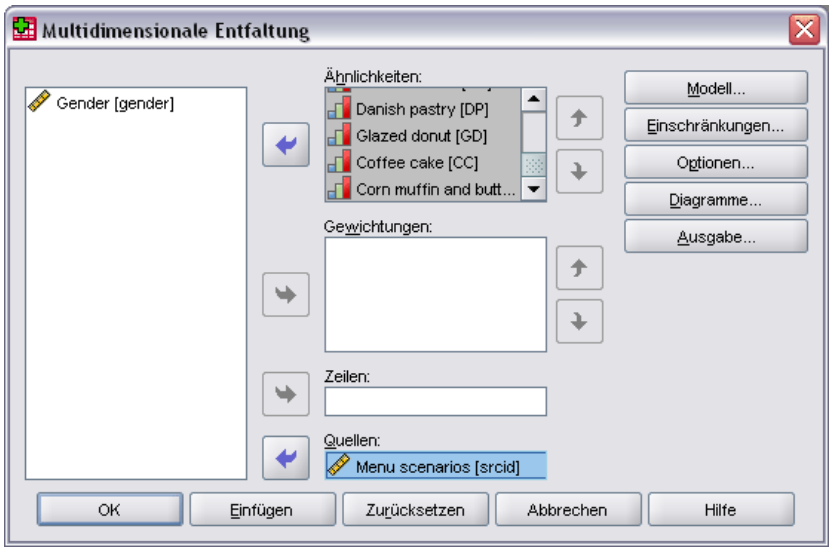
Annahmen. Es müssen mindestens zwei Variablen angegeben werden. Die Anzahl der Dimensionen in der Lösung darf die Anzahl der Objekte minus 1 nicht übersteigen. Wenn nur eine Quelle angegeben ist, sind alle Modelle äquivalent zum Identitätsmodell. In diesem Fall wird für die Analyse standardmäßig das Identitätsmodell verwendet.

So lassen Sie eine multidimensionale Entfaltung berechnen:

- Wählen Sie die folgenden Befehle aus den Menüs aus:

  Analysieren
  Skalieren
  Multidimensionale Entfaltung (PREFSCAL)...

Abbildung 8-1
Hauptdialogfeld "Multidimensionale Entfaltung"



- Wählen Sie zwei oder mehre Variablen aus, mit denen die Spalten in der rechtwinkligen Distanzmatrix identifiziert werden. Jede Variable stellt ein separates Spaltenobjekt dar.

- Wählen Sie optional Gewichtungsvariablen aus. Die Anzahl der Gewichtungsvariablen muss der Anzahl der ausgewählten Spaltenobjektvariablen entsprechen. Die Reihenfolge der Gewichtungsvariablen muss mit der Reihenfolge der Spaltenobjekte übereinstimmen, die sie gewichten.

- Wählen Sie optional eine Zeilenvariable aus. Die Werte (oder Wertelabels) dieser Variable werden verwendet, um die Zeilenobjekte in der Ausgabe zu beschriften.

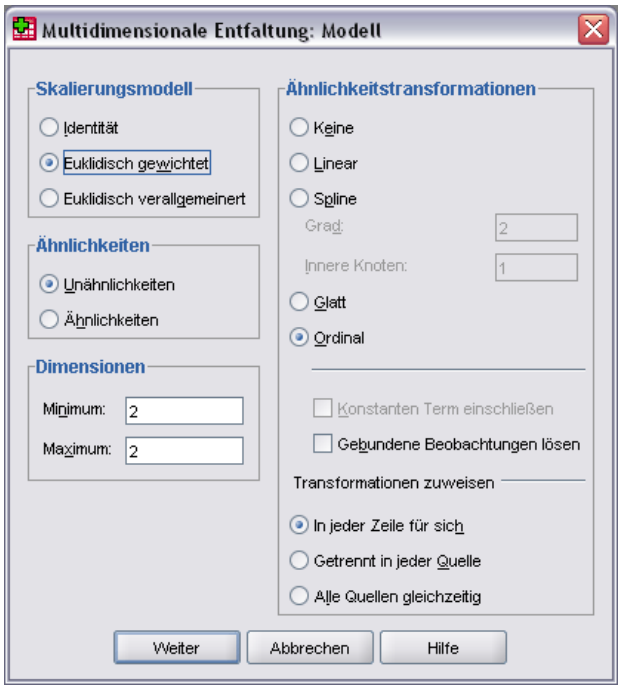
- Wenn mehrere Quellen vorhanden sind, wählen Sie bei Bedarf eine Quellenvariable aus. Die Anzahl der Fälle in der Datendatei muss der Anzahl der Zeilenobjekte multipliziert mit der Anzahl der Quellen entsprechen.

  Darüber hinaus können Sie ein Modell für die multidimensionale Entfaltung angeben, Einschränkungen für den gemeinsamen Raum festlegen, Konvergenzkriterien festlegen, die zu verwendende Ausgangskonfiguration angeben sowie Diagramme und die Ausgabe auswählen.

Definieren eines Modells für die multidimensionale Entfaltung

Im Dialogfeld "Modell" können Sie ein Skalierungsmodell, die minimale und maximale Anzahl seiner Dimensionen, die Struktur der Distanzmatrix und die für die Ähnlichkeiten zu verwendende Transformation angeben sowie festlegen, ob Ähnlichkeiten abhängig von der Zeile, abhängig von der Quelle oder unabhängig von der Quelle transformiert werden.

Abbildung 8-2
Dialogfeld "Modell"



Skalierungsmodell. Wählen Sie eine der folgenden Alternativen aus:

- **Identität.** Alle Quellen weisen dieselbe Konfiguration auf.
- **Euklidisch gewichtet.** Hierbei handelt es sich um ein Modell für individuelle Differenzen. Jede Quelle hat einen individuellen Raum, in dem jede Dimension des gemeinsamen Raums unterschiedlich gewichtet ist.
- **Euklidisch verallgemeinert.** Hierbei handelt es sich um ein Modell für individuelle Differenzen. Jede Quelle hat einen individuellen Raum, der einer Rotation des gemeinsamen Raums und einer darauf folgenden differenzierten Gewichtung der Dimensionen entspricht.

Ähnlichkeiten. Geben Sie an, ob die Distanzmatrix Ähnlichkeits- oder Unähnlichkeitsmaße enthält.

Dimensionen. Standardmäßig wird eine Lösung in zwei Dimensionen berechnet (Minimum = 2, Maximum = 2). Sie können für das Minimum und Maximum eine beliebige ganze Zahl zwischen 1 und der Anzahl der Objekte minus 1 angeben, solange das Minimum kleiner oder gleich dem Maximum ist. Die Prozedur berechnet eine Lösung in der höchsten Dimension und reduziert die Dimensionalität dann schrittweise, bis die niedrigste Dimensionalität erreicht ist.

Ähnlichkeitstransformationen. Wählen Sie eine der folgenden Alternativen aus:

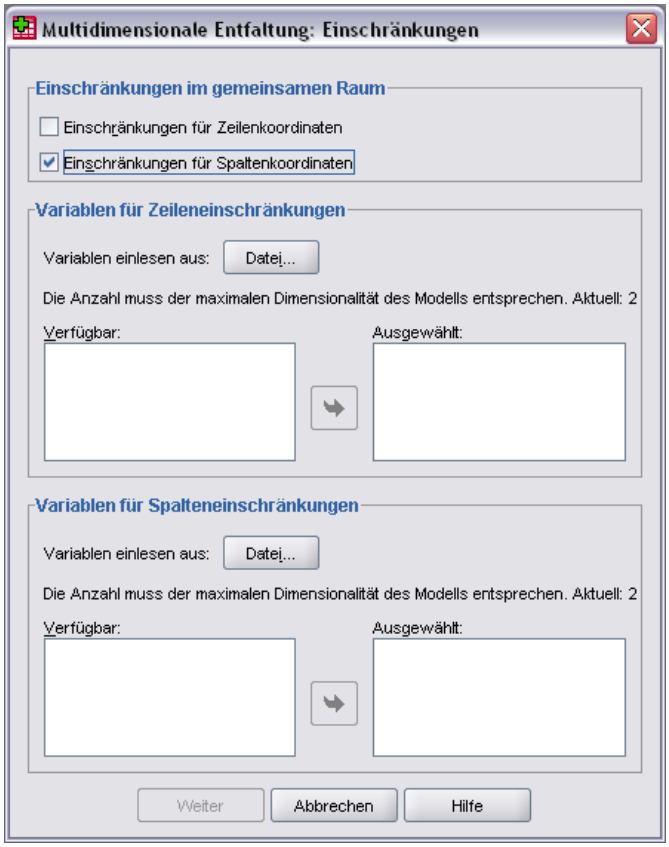
- **Keine.** Die Ähnlichkeiten werden nicht transformiert. Sie können optional Konstanten Term einschließen auswählen. In diesem Fall werden die Ähnlichkeiten durch einen konstanten Term verschoben.
- **Linear.** Die transformierten Ähnlichkeiten sind proportional zu den ursprünglichen Ähnlichkeiten. Hierbei wird in der Transformationsfunktion eine Steigung geschätzt und der Schnittpunkt mit der y-Achse auf 0 festgelegt. Dies wird auch als Verhältnistransformation bezeichnet. Sie können optional Konstanten Term einschließen auswählen. In diesem Fall werden die Ähnlichkeiten durch einen konstanten Term verschoben. Dies wird auch als Intervalltransformation bezeichnet.
- **Spline.** Die transformierten Ähnlichkeiten sind eine geglättete, nicht fallende, stückweise, polynomiale Transformation der ursprünglichen Ähnlichkeiten. Sie können den Grad des Polynoms und die Anzahl der inneren Knoten angeben. Sie können optional Konstanten Term einschließen auswählen. In diesem Fall werden die Ähnlichkeiten durch einen konstanten Term verschoben.
- **Glatt.** Die transformierten Ähnlichkeiten weisen dieselbe Ordnung wie die ursprünglichen Ähnlichkeiten auf. Es gilt jedoch eine Einschränkung, die die Unterschiede zwischen aufeinander folgenden Werten berücksichtigt. Das Ergebnis ist eine "glatte ordinale" Transformation. Sie können angeben, ob die Bindung von gebundenen Ähnlichkeiten erhalten bleiben muss oder aufgehoben werden darf.
- **Ordinal.** Die transformierten Ähnlichkeiten weisen dieselbe Ordnung wie die ursprünglichen Ähnlichkeiten auf. Sie können angeben, ob die Bindung von gebundenen Ähnlichkeiten erhalten bleiben muss oder aufgehoben werden darf.

Transformationen zuweisen. Geben Sie an, ob nur die Ähnlichkeiten innerhalb jeder Zeile oder nur die Ähnlichkeiten innerhalb jeder Quelle miteinander verglichen werden sollen oder ob die Vergleiche unabhängig von Zeile oder Quelle sind, also ob die Transformationen für einzelne Zeilen, für einzelne Quellen oder für alle Ähnlichkeiten gleichzeitig durchgeführt werden sollen.

Multidimensionale Entfaltung: Einschränkungen

Im Dialogfeld "Einschränkungen" können Sie Einschränkungen für den gemeinsamen Raum festlegen.

Abbildung 8-3
Dialogfeld "Einschränkungen"



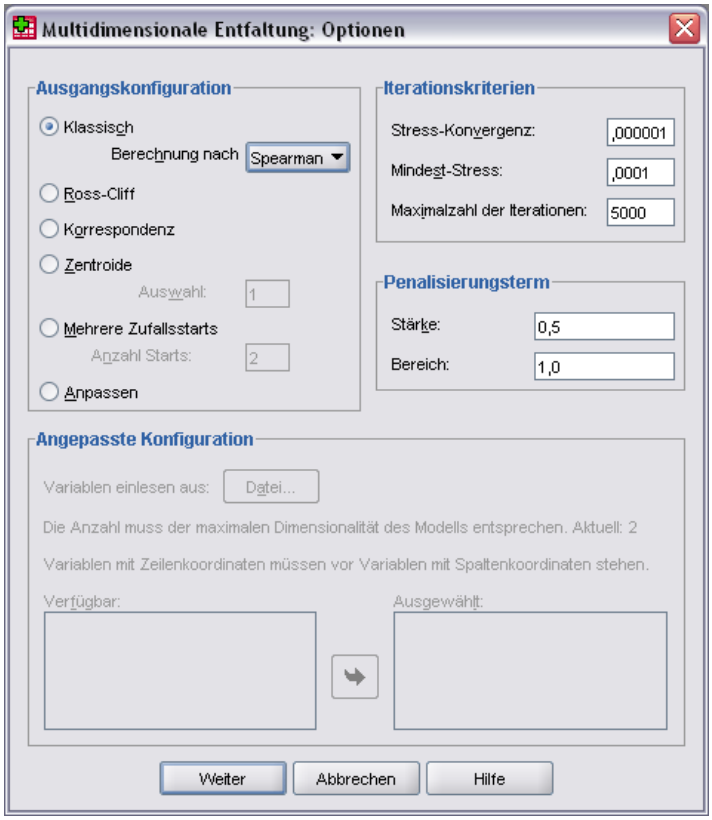
Einschränkungen im gemeinsamen Raum. Sie können die Koordinaten von Zeilen- und/oder Spaltenobjekten im gemeinsamen Raum festlegen.

Variablen für Zeilen-/Spalteneinschränkungen. Wählen Sie die Datei aus, die die Einschränkungen enthält, und wählen Sie die Variablen aus, die die Einschränkungen im gemeinsamen Raum definieren. Die erste ausgewählte Variable enthält die Koordinaten der Objekte in der ersten Dimension, die zweite Variable entspricht den Koordinaten der zweiten Dimension usw. Ein fehlender Wert zeigt an, dass eine Koordinate in einer Dimension frei ist. Die Anzahl der ausgewählten Variablen muss der angegebenen Höchstanzahl von Dimensionen entsprechen. Die Anzahl der Fälle jeder Variablen muss der Anzahl der Objekte entsprechen.

Multidimensionale Entfaltung: Optionen

Im Dialogfeld "Optionen" können Sie die Ausgangskonfiguration auswählen, Iterations- und Konvergenzkriterien angeben und den Penalisierungsterm für den Stress festlegen.

Abbildung 8-4
Dialogfeld "Optionen"



Ausgangskonfiguration. Wählen Sie eine der folgenden Optionen:

- **Klassisch.** Die rechtwinklige Distanzmatrix wird verwendet, um die Zwischenblöcke (Werte zwischen Zeilen und zwischen Spalten) der vollständig symmetrischen MDS-Matrix zu ergänzen. Sobald die vollständige Matrix gebildet ist, wird als Ausgangskonfiguration eine klassische Skalierungslösung verwendet. Die Zwischenblöcke können durch Ableiten von Werten mithilfe der Ungleichung im Dreieck oder der Spearman-Distanzen aufgefüllt werden.
- **Ross-Cliff.** Beim Ross-Cliff-Start werden die Ergebnisse einer Einzelwertzerlegung auf der zweifach zentrierten und quadratischen Distanzmatrix als Ausgangswerte für die Zeilen- und Spaltenobjekte verwendet.
- **Korrespondenz.** Beim Korrespondenzstart werden die Ergebnisse einer Korrespondenzanalyse der Daten in umgekehrter Reihenfolge (Ähnlichkeiten anstelle von Unähnlichkeiten) mit einer symmetrischen Normalisierung von Zeilen- und Spaltenwerten verwendet.
- **Zentroide.** Die Prozedur beginnt mit der Positionierung von Zeilenobjekten in der Konfiguration auf der Grundlage einer Eigenwertzerlegung. Anschließend werden die Spaltenobjekte auf dem Zentroid der angegebenen Auswahl positioniert. Geben Sie für die Anzahl der Auswahl eine positive ganze Zahl zwischen 1 und der Anzahl der Ähnlichkeitsvariablen an.

- **Mehrere Zufallsstarts.** Die Lösungen werden für verschiedene zufällig ausgewählte Ausgangskonfigurationen berechnet. Die Lösung mit dem kleinsten penalisierten Stress wird als beste Lösung angezeigt.
- **Anpassen.** Sie können Variablen auswählen, die die Koordinaten Ihrer eigenen Ausgangskonfiguration enthalten. Die Anzahl der ausgewählten Variablen muss der angegebenen Höchstanzahl von Dimensionen entsprechen, wobei die erste Variable den Koordinaten von Dimension 1, die zweite Variable den Koordinaten von Dimension 2 entspricht usw. Die Anzahl der Fälle in jeder Variablen muss der Gesamtzahl der Zeilen- und Spaltenobjekte entsprechen. Die Zeilen- und Spaltenkoordinaten müssen gestapelt sein, wobei die Spaltenkoordinaten den Zeilenkoordinaten folgen.

Iterationskriterien. Geben Sie Werte für Iterationskriterien an.

- **Stress-Konvergenz.** Der Algorithmus beendet die Iterationen, wenn die relative Differenz von aufeinander folgenden penalisierten Stresswerten unter dem hier angegebenen Wert liegt. Dieser Wert darf nicht negativ sein.
- **Mindest-Stress.** Der Algorithmus wird beendet, wenn der penalisierte Stress unter den hier angegebenen Wert fällt. Dieser Wert darf nicht negativ sein.
- **Maximalzahl der Iterationen.** Der Algorithmus führt die hier angegebene Anzahl von Iterationen durch, falls nicht vorher eines der oben genannten Kriterien erfüllt wird.

Penalisierungsterm. Der Algorithmus versucht, den penalisierten Stress zu minimieren. Der penalisierte Stress ist ein Maß für die Güte der Anpassung, das dem Produkt des Kruskal-Stress-I und einem Penalisierungsterm auf der Basis des Variationskoeffizienten der transformierten Ähnlichkeiten entspricht. In diesen Feldern können Sie die Stärke und den Bereich des Penalisierungsterms festlegen.

- **Stärke.** Je kleiner der Wert des Stärkeparameters ist, desto stärker wird die Penalisierung. Geben Sie einen Wert zwischen 0,0 und 1,0 an.
- **Bereich.** Mit diesem Parameter legen Sie den Moment fest, ab dem die Penalisierung erfolgt. Bei 0,0 erfolgt keine Penalisierung. Höhere Wert führen dazu, dass der Algorithmus eine Lösung mit einer größeren Variation innerhalb der transformierten Ähnlichkeiten sucht. Geben Sie einen nichtnegativen Wert an.

Multidimensionale Entfaltung: Diagramme

Im Dialogfeld “Diagramme” können Sie die zu erstellenden Diagramme festlegen.

Abbildung 8-5
Dialogfeld "Diagramme"

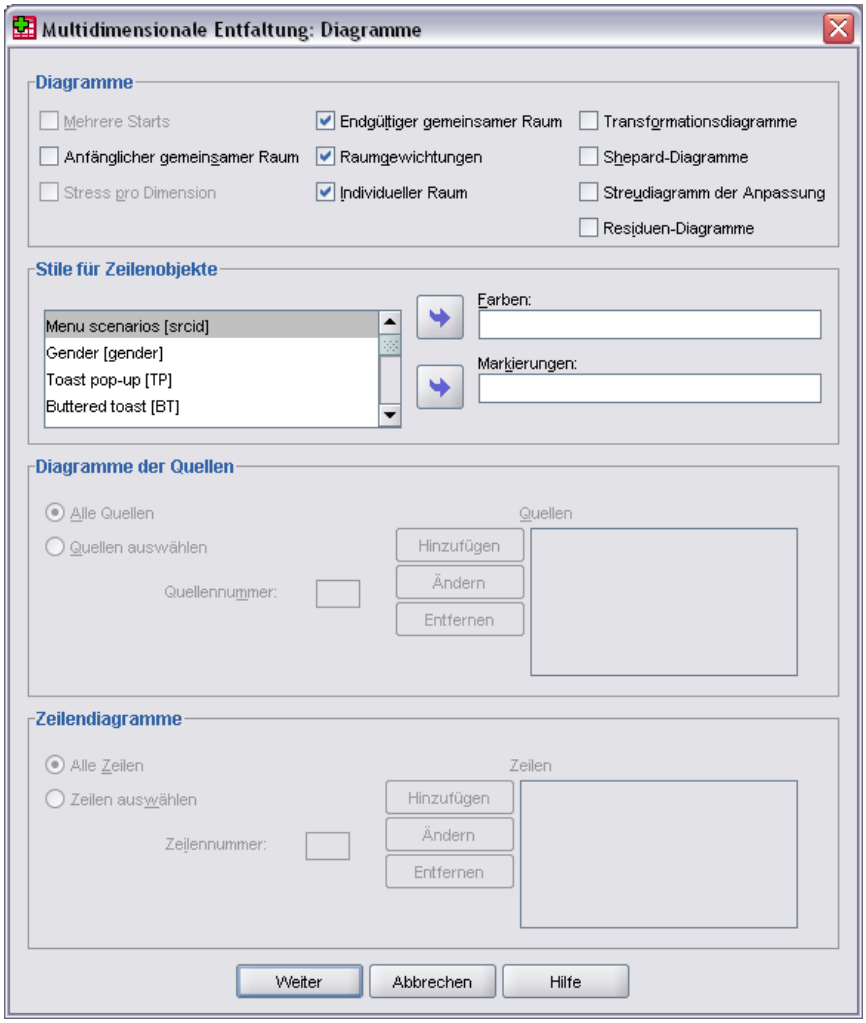


Diagramme. Folgende Diagramme sind verfügbar:

- **Mehrere Starts.** Zeigt ein gestapeltes Histogramm des penalisierten Stress mit Stress und Penalisierung an.
- **Anfänglicher gemeinsamer Raum.** Zeigt eine Streudiagramm-Matrix der Koordinaten des anfänglichen gemeinsamen Raums an.
- **Stress pro Dimension.** Erstellt ein Liniendiagramm des penalisierten Stress im Vergleich zur Dimensionalität. Dieses Diagramm wird nur erstellt, wenn die Höchstanzahl von Dimensionen größer als die Mindestanzahl von Dimensionen ist.
- **Endgültiger gemeinsamer Raum.** Eine Streudiagramm-Matrix der Koordinaten des gemeinsamen Raums wird angezeigt.
- **Raumgewichtungen.** Ein Streudiagramm der individuellen Raumgewichtungen wird erstellt. Dies ist nur möglich, wenn im Dialogfeld "Modell" eines der Modelle für individuelle Differenzen angegeben ist. Für das Modell "Euklidisch gewichtet" werden die Gewichtungen

für alle Quellen mit einer Dimension auf jeder Achse in einem Diagramm angezeigt. Für das Modell "Euklidisch verallgemeinert" wird ein Diagramm pro Dimension erstellt, das für jede Quelle die Rotation und Gewichtung der jeweiligen Dimension anzeigt.

- **Individueller Raum.** Eine Streudiagramm-Matrix der Koordinaten des individuellen Raums jeder Quelle wird angezeigt. Dies ist nur möglich, wenn im Dialogfeld "Modell" eines der Modelle für individuelle Differenzen angegeben ist.
- **Transformationsdiagramme.** Es wird ein Streudiagramm der ursprünglichen Ähnlichkeiten im Vergleich zu den transformierten Ähnlichkeiten erstellt. Abhängig davon, wie Transformationen angewendet werden, wird jeder Zeile oder Quelle eine separate Farbe zugewiesen. Eine unabhängige Transformation erzeugt eine Farbe.
- **Shepard-Diagramme.** Die ursprünglichen Ähnlichkeiten im Vergleich zu sowohl transformierten Ähnlichkeiten als auch Distanzen. Die Distanzen werden durch Punkte angezeigt, die transformierten Ähnlichkeiten durch eine Linie. Abhängig davon, wie Transformationen angewendet werden, wird für jede Zeile oder Quelle eine separate Linie gezeichnet. Eine unabhängige Transformation erzeugt eine Linie.
- **Streudiagramm der Anpassung.** Es wird ein Streudiagramm der transformierten Ähnlichkeiten im Vergleich zu den Distanzen angezeigt. Wenn mehrere Quellen angegeben werden, erhält jede Quelle eine separate Farbe.
- **Residuen-Diagramme.** Es wird ein Streudiagramm der transformierten Ähnlichkeiten im Vergleich zu den Residuen (transformierte Ähnlichkeiten minus Distanzen) angezeigt. Wenn mehrere Quellen angegeben werden, erhält jede Quelle eine separate Farbe.

Stile für Zeilenobjekte. Hiermit können Sie umfassender festlegen, wie Zeilenobjekte in den Diagrammen dargestellt werden. Die Werte der optionalen Farbvariablen werden verwendet, um alle Farben zu durchlaufen. Die Werte der optionalen Markierungsvariablen werden verwendet, um alle möglichen Markierungen zu durchlaufen.

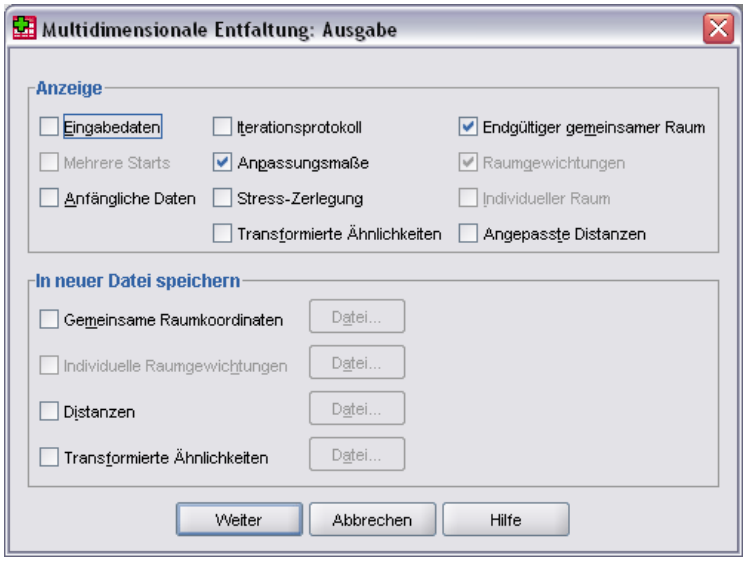
Diagramme der Quellen. Für Individueller Raum, Streudiagramm der Anpassung und Residuen-Diagramme (sowie, falls Transformationen nach Quellen angewendet werden, für Transformationsdiagramme und Shepard-Diagramme) können Sie angeben, für welche Quellen Diagramme erstellt werden sollen. Die eingegebenen Quellennummern müssen Werte der im Hauptdialogfeld angegebenen Quellenvariable sein und im Bereich von 1 bis zur Anzahl der Quellen liegen.

Zeilendiagramme. Wenn Transformationen für Zeilen erfolgen, können Sie für Transformationsdiagramme und Shepard-Diagramme die Zeile angeben, für die die Diagramme erstellt werden sollen. Die eingegebenen Zeilenanzahl muss im Bereich von 1 bis zur Anzahl der Zeilen liegen.

Multidimensionale Entfaltung: Ausgabe

Im Dialogfeld "Ausgabe" können Sie die angezeigte Ausgabe steuern und Auszüge davon in eigenen Dateien speichern.

Abbildung 8-6
Dialogfeld "Ausgabe"



Anzeige. Wählen Sie mindestens eine der folgenden Anzeigeoptionen aus:

- **Eingabedaten.** Zeigt die ursprünglichen Ähnlichkeiten und, sofern vorhanden, die Datengewichtungen, die Ausgangskonfiguration und die festen Koordinaten an.
- **Mehrere Starts.** Zeigt den Startwert für Zufallszahlen und den penalisierten Stresswert für jeden Zufallsstart an.
- **Anfängliche Daten.** Zeigt die Koordinaten des anfänglichen gemeinsamen Raums an.
- **Iterationsprotokoll.** Zeigt das Iterationsprotokoll des Hauptalgorithmus an.
- **Anpassungsmaße.** Zeigt verschiedene Maße an. Die Tabelle enthält mehrere Maße für die Güte bzw. mangelhafte Güte der Anpassung, Korrelation, Variation und Nichtdegenerierung.
- **Stress-Zerlegung.** Zeigt die Zerlegung des penalisierten Stress von Objekten, Zeilen und Quellen an, einschließlich der Mittelwerte und Standardabweichungen von Zeilen, Spalten und Quellen.
- **Transformierte Ähnlichkeiten.** Zeigt die transformierten Ähnlichkeiten an.
- **Endgültiger gemeinsamer Raum.** Zeigt die Koordinaten des gemeinsamen Raums an.
- **Raumgewichtungen.** Zeigt die individuellen Raumgewichtungen an. Diese Option ist nur verfügbar, wenn eines der Modelle für individuelle Differenzen angegeben wird. Je nach Modell werden die Raumgewichtungen in Rotations- und Dimensionsgewichtungen zerlegt, die ebenfalls angezeigt werden.
- **Individueller Raum.** Zeigt die Koordinaten der individuellen Räume an. Diese Option ist nur verfügbar, wenn eines der Modelle für individuelle Differenzen angegeben wird.
- **Angepasste Distanzen.** Zeigt die Distanzen zwischen den Objekten in der Konfiguration an.

In neuer Datei speichern. Sie können die Koordinaten des gemeinsamen Raums, die individuellen Raumgewichtungen, die Distanzen und die transformierten Ähnlichkeiten in eigenen SPSS-Datendateien speichern.

Zusätzliche Funktionen beim Befehl PREFSCAL

Sie können die multidimensionale Entfaltung der Ähnlichkeiten individuell anpassen, indem Sie die Auswahl in ein Syntaxfenster einfügen und die daraus resultierende Befehlssyntax für PREFSCAL bearbeiten. Mit der Befehlssyntax-Sprache verfügen Sie außerdem über folgende Möglichkeiten:

- Wenn mehrere Quellen verfügbar sind, können Sie für individuelle Räume, Streudiagramme der Anpassung und Residuendiagramme – bei Transformationen von Matrizen auch für Transformationsdiagramme und Shepard-Diagramme – Listen mit mehreren Quellen angeben (mit dem Unterbefehl PLOT).
- Bei Transformationen von Zeilen können Sie für Transformationsdiagramme und Shepard-Diagramme Listen mit mehreren Zeilen angeben (mit dem Unterbefehl PLOT).
- Sie können anstelle einer Zeilen-ID-Variablen eine Anzahl von Zeilen angeben (mit dem Unterbefehl INPUT).
- Sie können anstelle einer Quellen-ID-Variablen eine Anzahl von Quellen angeben (mit dem Unterbefehl INPUT).

Vollständige Informationen zur Syntax finden Sie in der *Command Syntax Reference*.

Teil II: Beispiele

Kapitel
9

Kategoriale Regression

Das Ziel der kategorialen Regression mit optimaler Skalierung liegt darin, die Beziehung zwischen einer Antwortvariablen und einer Gruppe von Einflußgrößen zu beschreiben. Durch die Quantifizierung dieser Beziehung können Werte der Antwort für jede Kombination von Einflußgrößen vorhergesagt werden.

In diesem Kapitel werden die Analysen im Rahmen der Regression mit optimaler Skalierung anhand von zwei Beispielen dargestellt. Das erste Beispiel erläutert die grundlegenden Konzepte mithilfe eines kleinen Daten-Sets. Im zweiten Beispiel wird eine deutlich größere Gruppe von Variablen und Beobachtungen herangezogen.

Beispiel: Daten für Teppichreiniger

In einem beliebten Beispiel möchte die Firma (Green als auch Wind, 1973) einen neuen Teppichreiniger vermarkten und dazu den Einfluss von fünf Faktoren auf die Bevorzugung durch den Verbraucher untersuchen: Verpackungsgestaltung, Markenname, Preis, Gütesiegel *Good Housekeeping* und Geld-zurück-Garantie. Die Verpackungsgestaltung liegt in drei Faktorstufen vor, die sich durch die Position der Auftragebürste unterscheiden. Außerdem gibt es drei Markennamen (*K2R*, *Glory* und *Bissell*), drei Preisstufen sowie je zwei Stufen (Nein oder Ja) für die letzten beiden Faktoren. Die nachstehende Tabelle zeigt die Variablen in der Teppichreiniger-Untersuchung mit den zugehörigen Beschriftungen und Werten.

Tabelle 9-1
Erklärende Variablen in der Teppichreiniger-Untersuchung

Variablenname	Variablenlabel	Wertelabels
package	Package design	A*, B*, C*
brand	Brand name	K2R, Glory, Bissell
price	Price	$1.19, $1.39, $1.59
seal	Gütesiegel *Good Housekeeping*	Nein, Ja
money	Money-back guarantee	Nein, Ja

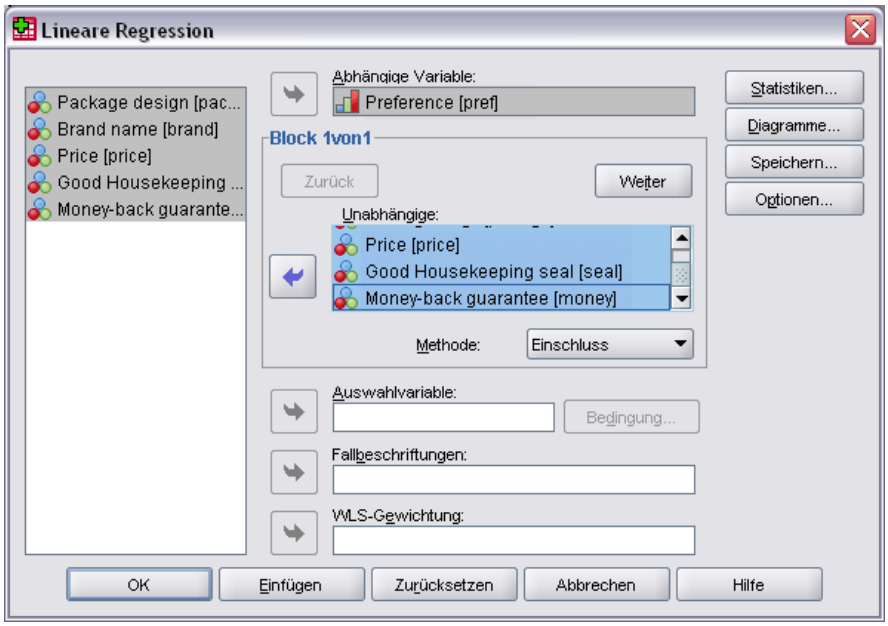
10 Kunden stufen 22 Profile ein, die durch diese Faktoren definiert sind. Die Variable *Preference* enthält den Rang der durchschnittlichen Einstufung für die verschiedenen Profile. Ein niedriger Rang bedeutet eine starke Bevorzugung. Diese Variable gibt ein Gesamtmaß der Bevorzugung für die Profile an. Mithilfe der kategorialen Regression untersuchen Sie, wie die fünf Faktoren mit der Bevorzugung zusammenhängen. Dieses Daten-Set befindet sich in *carpet.sav*. Für weitere Informationen siehe Beispieldateien in Anhang A auf S. 326.

Standardverfahren der linearen Regressionsanalyse

- Zum Erzeugen der Ausgabe nach dem Standardverfahren der linearen Regression wählen Sie die folgenden Befehle aus den Menüs aus:

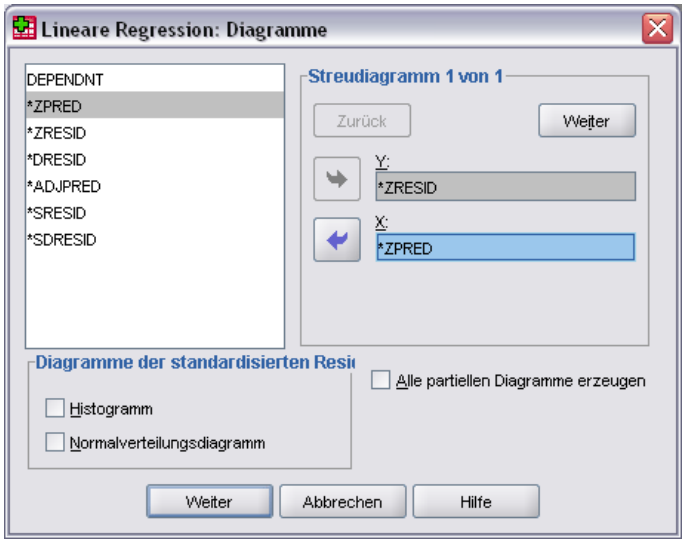
  Analysieren
  Regression
  Linear...

Abbildung 9-1
Dialogfeld "Lineare Regression"



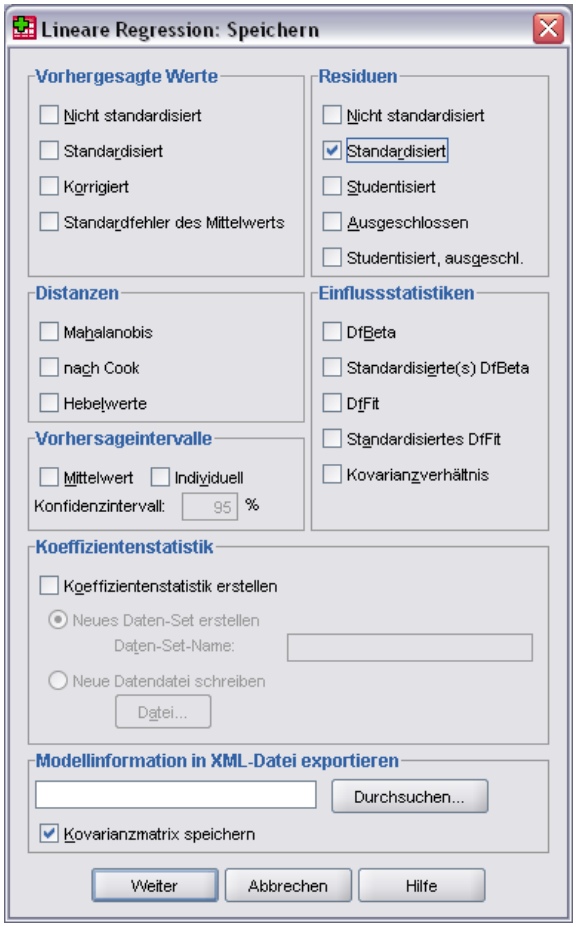
- Wählen Sie *Preference* als abhängige Variable aus.
- Wählen Sie die Einträge *Package design* bis *Money-back guarantee* als unabhängige Variablen aus.
- Klicken Sie auf Diagramme.

Abbildung 9-2
Dialogfeld "Diagramme"



- Wählen Sie **ZRESID* als *y*-Variable aus.
- Wählen Sie **ZPRED* als *x*-Variable aus.
- Klicken Sie auf Weiter.
- Klicken Sie im Dialogfeld "Lineare Regression" auf Speichern.

Abbildung 9-3
Dialogfeld "Speichern"



- Wählen Sie in der Gruppe "Residuen" die Option Standardisiert.
- Klicken Sie auf Weiter.
- Klicken Sie im Dialogfeld "Lineare Regression" auf OK.

Modellzusammenfassung

Abbildung 9-4
Modellzusammenfassung für lineare Regression nach dem Standardverfahren

Modellzusammenfassung[b]

Modell	R	R-Quadrat	Korrigiertes R-Quadrat	Standardfehler des Schätzers
1	,841[a]	,707	,615	3,99810

a. Einflußvariablen : (Konstante), Good Housekeeping seal, Price, Money-back guarantee, Brand name, Package design

b. Abhängige Variable: Preference

Die lineare Regression ist das Standardverfahren zum Beschreiben der Beziehungen in diesem Problem. Das am weitesten verbreitete Maß für die Anpassungsgüte des Regressionsmodells ist R^2. Diese Statistik zeigt, inwieweit die Varianz in der Antwort durch die gewichtete Kombination der Einflussgrößen erläutert wird. Je dichter R^2 an 1 liegt, desto höher ist die Anpassungsgüte des Modells. Die Regression von *Preference* auf die fünf Einflußgrößen führt zu einem Wert von 0,707 für R^2. Dies bedeutet, dass etwa 71 % der Varianz in der Einstufung der Bevorzugung durch die Einflußgrößen in der linearen Regression erklärt werden.

Koeffizienten

Die standardisierten Koeffizienten sind in der Tabelle aufgeführt. Das Vorzeichen der Koeffizienten zeigt, ob die vorhergesagte Antwort ansteigt oder abfällt, wenn die Einflußgröße wächst und alle anderen Einflußgrößen konstant bleiben. Bei kategorialen Daten geht die Bedeutung eines Anstiegs bei einer Einflußgröße aus der Kodierung der Kategorien hervor. Ein Anstieg bei *Money-back guarantee*, *Package design* oder *Good Housekeeping seal* führt beispielsweise zu einer niedrigeren vorhergesagten Einstufung der Bevorzugung. *Money-back guarantee* erhält die Kodierung 1 für *no money-back guarantee* bzw. die Kodierung 2 für *money-back guarantee*. Ein Anstieg bei *Money-back guarantee* bedeutet, dass eine Geld-zurück-Garantie hinzugefügt wurde. Die Einführung einer Geld-zurück-Garantie senkt daher die vorhergesagte Einstufung der Bevorzugung, bewirkt also eine höhere vorhergesagte Bevorzugung.

Abbildung 9-5
Regressionskoeffizienten

Koeffizienten[a]

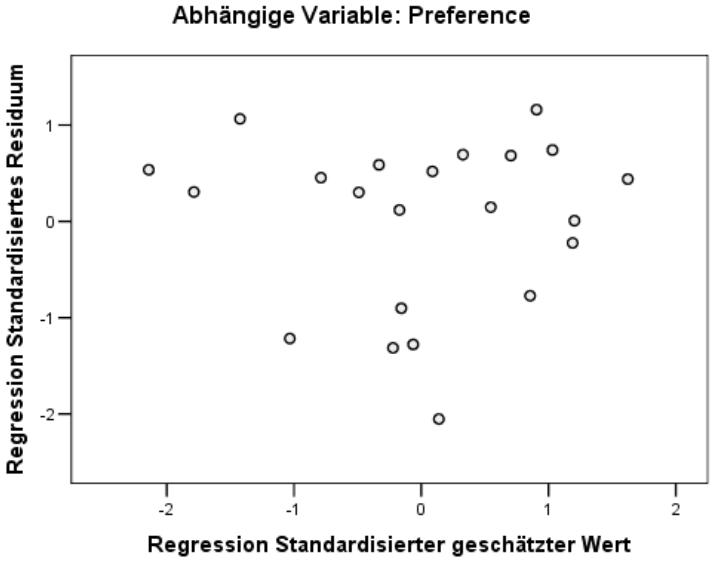
Modell		Nicht standardisierte Koeffizienten		Standardisierte Koeffizienten		
		B	Standardfehler	Beta	T	Signifikanz
1	(Konstante)	22,529	5,177		4,352	,000
	Package design	-4,159	1,036	-,560	-4,015	,001
	Brand name	,429	1,054	,056	,407	,689
	Price	2,703	1,009	,366	2,681	,016
	Money-back guarantee	-2,779	1,921	-,197	-1,447	,167
	Good Housekeeping seal	-4,314	1,780	-,330	-2,423	,028

a. Abhängige Variable: Preference

Der Wert des Koeffizienten gibt das Maß für die Änderung der vorhergesagten Einstufung der Bevorzugung an. Mithilfe standardisierter Koeffizienten werden die Interpretationen auf der Grundlage der Standardabweichungen der Variablen vorgenommen. Jeder Koeffizient zeigt die Anzahl der Standardabweichungen, um die sich die vorhergesagte Antwort pro Änderung in der Standardabweichung in einer Einflußgröße ändert, wenn alle anderen Einflußgrößen konstant bleiben. Eine Änderung um eine Standardabweichung in *Brand name* führt zu einer höheren vorhergesagten Bevorzugung von 0,056 Standardabweichungen. Die Standardabweichung für *Preference* ist 6,44; *Preference* steigt somit um $0,056 \times 6,44 = 0,361$. Änderungen an *Package design* bewirken die größten Änderungen bei der vorhergesagten Bevorzugung.

Residuen-Streudiagramme

Abbildung 9-6
Residuen im Vergleich zu vorhergesagten Werten

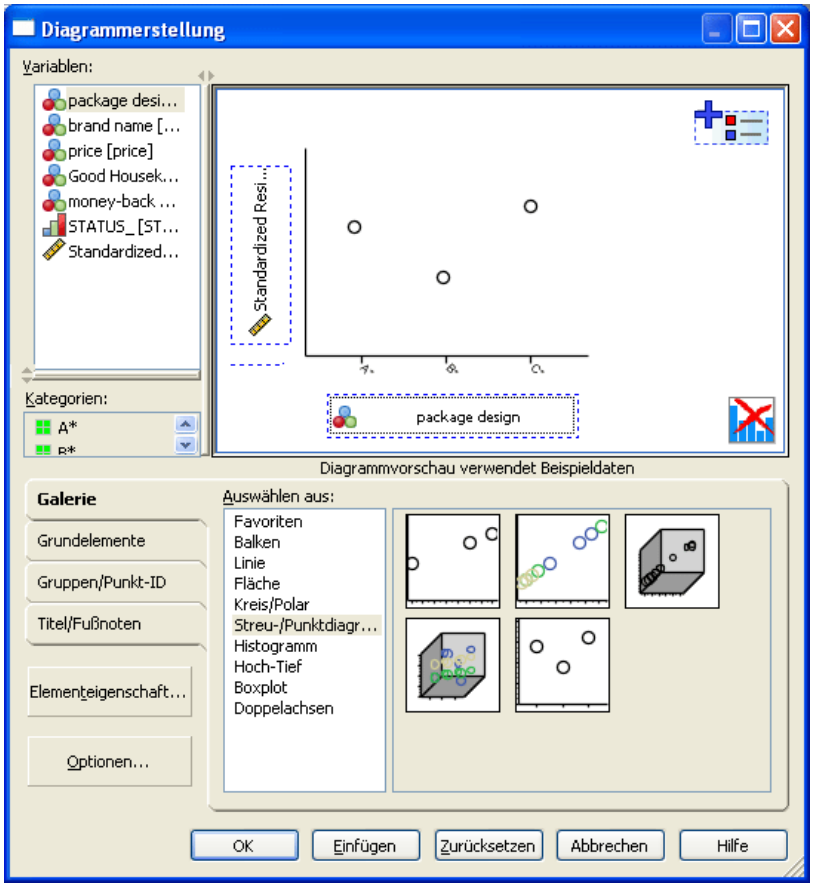


Die standardisierten Residuen werden anhand der standardisierten vorhergesagten Werte geplottet. Bei einer hohen Anpassungsgüte des Modells sollten keine Muster entstehen. Hier ist eine U-Form sichtbar, bei der sowohl hohe als auch niedrige standardisierte vorhergesagte Werte positive Residuen besitzen. Die standardisierten vorhergesagten Werte nahe 0 weisen eher negative Werte auf.

► Zum Erstellen eines Streudiagramms der Residuen nach der Einflußgröße *Package design* wählen Sie die folgenden Befehle aus den Menüs aus:

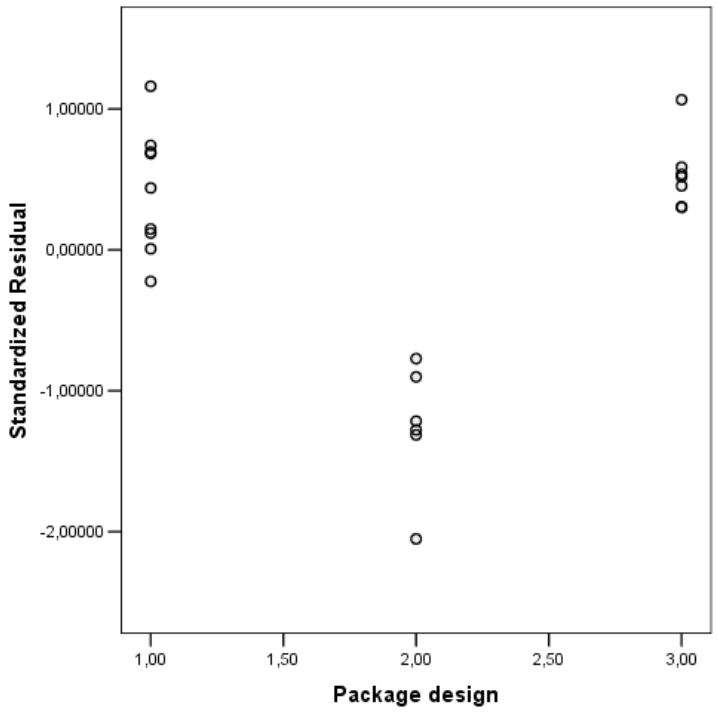
Grafiken
Diagrammerstellung...

Abbildung 9-7
Diagrammerstellung



- Wählen Sie die Galerie Streu-/Punktdiagramm aus und wählen Sie die Option “Einfaches Streudiagramm”.
- Wählen Sie *Standardisiertes Residuum* als *y*-Achsen-Variable aus und *Package design* als *x*-Achsen-Variable.
- Klicken Sie auf OK.

Abbildung 9-8
Residuen im Vergleich mit der Verpackungsgestaltung



Die U-Form ist im Diagramm der standardisierten Residuen im Vergleich mit der Verpackung stärker ausgeprägt. Alle Residuen für Design B* sind negativ; bei den anderen beiden Designs sind dagegen alle Residuen (bis auf eines) positiv. Beim linearen Regressionsmodell wird je ein Parameter für die einzelnen Variablen herangezogen. Die Beziehung kann daher nicht über das Standardverfahren festgehalten werden.

Kategoriale Regressionsanalyse

Der kategoriale Charakter der Variablen und die nichtlineare Beziehung zwischen *Preference* und *Package design* legen nahe, dass die Regression mit optimalen Werten bessere Ergebnisse liefern würde als die Regression im Standardverfahren. Die U-Form der Residuendiagramme weist darauf hin, dass *Package design* nominal behandelt werden sollte. Alle anderen Einflußgrößen werden auf dem entsprechenden numerischen Skalierungsniveau behandelt.

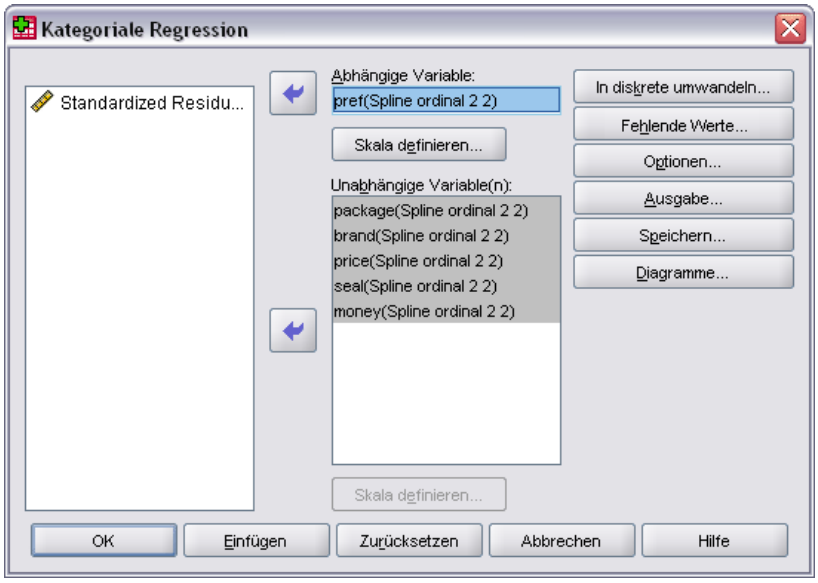
Die Antwortvariable verdient besondere Aufmerksamkeit. Es sollen die Werte für *Preference* vorhergesagt werden. Aus diesem Wert sollten so viele Eigenschaften der zugehörigen Kategorien wie möglich in den Quantifikationen eingebracht werden. Bei einem ordinalen oder nominalen Messniveau für die Skalierung werden die Differenzen zwischen den Antwortkategorien ignoriert. Werden die Antwortkategorien dagegen linear transformiert, bleiben die Differenzen zwischen den Kategorien erhalten. Die Antwort sollte somit numerisch skaliert werden; diese Vorgehensweise wird hier verfolgt.

Durchführung der Analyse

▶ Zum Ausführen einer kategorialen Regressionsanalyse wählen Sie die folgenden Befehle aus den Menüs aus:

Analysieren
  Regression
    Optimale Skalierung...

Abbildung 9-9
Dialogfeld "Kategoriale Regression"

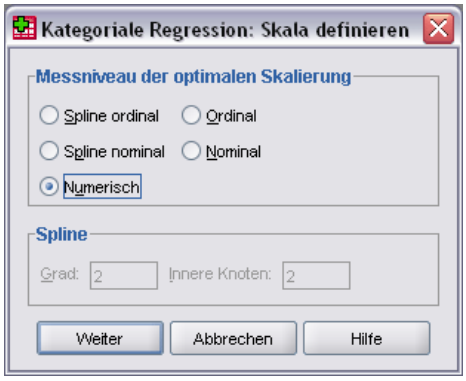


▶ Wählen Sie *Preference* als abhängige Variable aus.

▶ Wählen Sie die Einträge *Package design* bis *Money-back guarantee* als unabhängige Variablen aus.

▶ Wählen Sie *Preference* aus und klicken Sie auf Skala definieren.

Abbildung 9-10
Dialogfeld "Skala definieren"

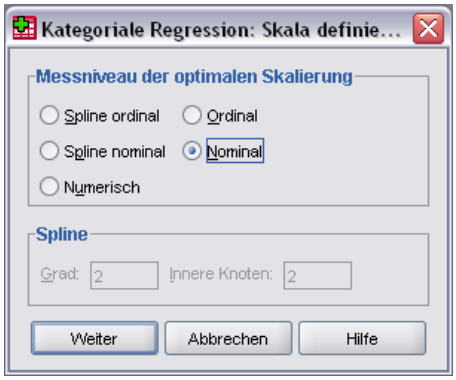


▶ Wählen Sie Numerisch als optimales Messniveau für die Skalierung.

▶ Klicken Sie auf Weiter.

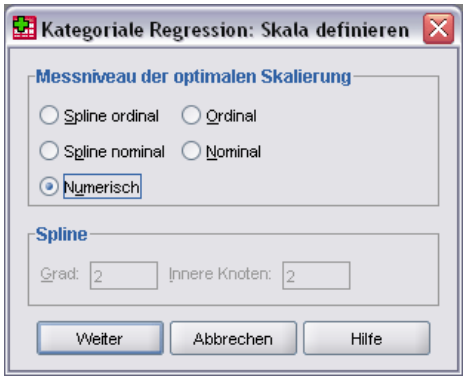
- Wählen Sie *Package design* aus und klicken Sie im Dialogfeld "Kategoriale Regression" auf Skala definieren.

Abbildung 9-11
Dialogfeld "Skala definieren"



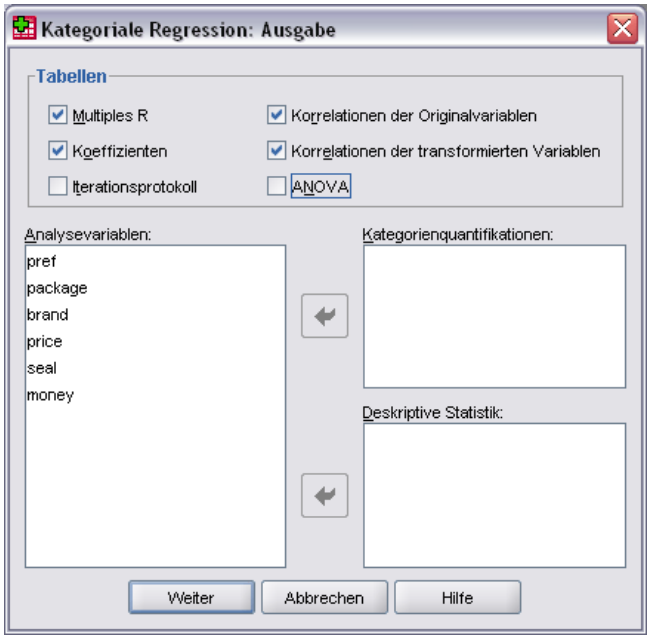
- Wählen Sie Nominal als optimales Messniveau für die Skalierung.
- Klicken Sie auf Weiter.
- Wählen Sie *Brand name* bis *Money-back guarantee* aus und klicken Sie im Dialogfeld "Kategoriale Regression" auf Skala definieren.

Abbildung 9-12
Dialogfeld "Skala definieren"



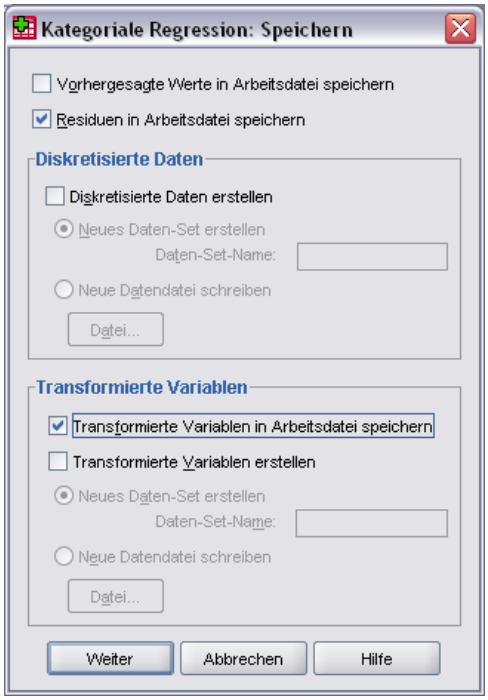
- Wählen Sie Numerisch als optimales Messniveau für die Skalierung.
- Klicken Sie auf Weiter.
- Klicken Sie im Dialogfeld "Kategoriale Regression" auf Ausgabe.

Abbildung 9-13
Dialogfeld "Ausgabe"



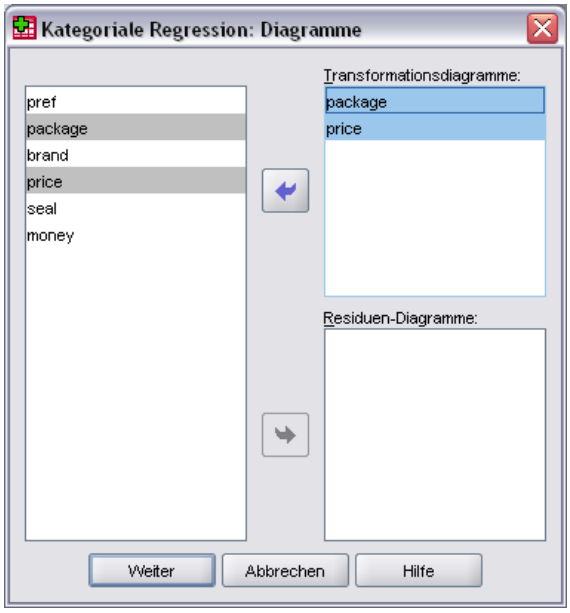
- Wählen Sie Korrelationen der Originalvariablen und Korrelationen der transformierten Variablen.
- Deaktivieren Sie die Option ANOVA.
- Klicken Sie auf Weiter.
- Klicken Sie im Dialogfeld "Kategoriale Regression" auf Speichern.

Abbildung 9-14
Dialogfeld "Speichern"



- Wählen Sie die Option Residuen in Arbeitsdatei speichern aus.
- Wählen Sie Im Gruppenfeld "Transformierte Variablen" die Option Transformierte Variablen in Arbeitsdatei speichern aus.
- Klicken Sie auf Weiter.
- Klicken Sie im Dialogfeld "Kategoriale Regression" auf Diagramme.

Abbildung 9-15
Dialogfeld "Diagramme"



- Legen Sie fest, dass Transformationsdiagramme für *Package design* und *Price* angelegt werden sollen.
- Klicken Sie auf Weiter.
- Klicken Sie im Dialogfeld "Kategoriale Regression" auf OK.

Interkorrelationen

Über die Interkorrelationen zwischen den Einflußgrößen wird die Multi-Kollinearität in der Regression ersichtlich. Hoch korrelierte Variablen führen zu instabilen Regressionsschätzungen. Durch die hohe Korrelation würde das Auslassen einer dieser Variablen aus dem Modell die Vorhersage jedoch nur minimal beeinflussen. Die Varianz in der Antwort, die sich aus der ausgelassenen Variablen ergibt, kann durch die verbleibende korrelierte Variable erklärt werden. Korrelationen nullter Ordnung sind allerdings empfindlich gegenüber Ausreißern. Auch die Multi-Kollinearität aufgrund einer hohen Korrelation zwischen einer Einflußgröße und einer Kombination anderer Einflußgrößen kann nicht erkannt werden.

Abbildung 9-16
Korrelationen der ursprünglichen Einflußgrößen

	Package design	Brand name	Price	Good Housekeeping seal	Money-back guarantee
Package design	1,000	-,189	-,126	,081	,066
Brand name	-,189	1,000	,065	-,042	-,034
Price	-,126	,065	1,000	,000	,000
Good Housekeeping seal	,081	-,042	,000	1,000	-,039
Money-back guarantee	,066	-,034	,000	-,039	1,000
Dimension	1	2	3	4	5
Eigenwert	1,291	1,038	,980	,905	,785

Abbildung 9-17
Korrelationen der transformierten Einflußgrößen

	Package design	Brand name	Price	Good Housekeeping seal	Money-back guarantee
Package design	1,000	-,156	-,089	,032	,102
Brand name	-,156	1,000	,065	-,042	-,034
Price	-,089	,065	1,000	,000	,000
Good Housekeeping seal	,032	-,042	,000	1,000	-,039
Money-back guarantee	,102	-,034	,000	-,039	1,000
Dimension	1	2	3	4	5
Eigenwert	1,248	1,043	,983	,905	,821

Die Interkorrelationen der Einflußgrößen für die nicht transfomierten und die transfomierten Einflußgrößen werden dargestellt. Alle Werte liegen nahe 0; dies bedeutet, dass Multi-Kollinearität zwischen einzelnen Variablen kein Problem ist.

Beachten Sie, dass geänderte Korrelationen nur bei *Package design* auftreten. Alle anderen Einflußgrößen werden numerisch behandelt; die Differenzen zwischen den Kategorien und die Reihenfolge der Kategorien bleiben für diese Variablen also erhalten. Das bedeutet wiederum, dass die Korrelationen sich nicht ändern.

Anpassungsgüte des Modells und Koeffizienten

Die Prozedur "Kategoriale Regression" liefert einen Wert von 0,948 für R^2. Dies bedeutet, dass nahezu 95 % der Varianz in der Einstufung der transformierten Bevorzugung durch die Regression der optimal transformierten Einflußgrößen erklärt werden. Durch Transformieren der Einflußgrößen erzielen Sie eine höhere Anpassungsgüte im Vergleich zum Standardverfahren.

Abbildung 9-18
Modellzusammenfassung für kategoriale Regression

Multiples R	R-Quadrat	Korrigiertes R-Quadrat
,974	,948	,927

Abhängige Variable: Preference
Einflußvariablen: Package design Brand name Price Good Housekeeping seal Money-back guarantee

Die nachstehende Tabelle zeigt die standardisierten Regressionskoeffizienten. Bei der kategorialen Regression werden die Variablen standardisiert; aus diesem Grund werden nur standardisierte Koeffizienten zurückgegeben. Diese Werte werden durch die zugehörigen Standardfehler dividiert und ergeben so jeweils einen *F*-Test für die einzelnen Variablen. Der Test für jede Variable ist allerdings davon abhängig, dass die anderen Einflußgrößen im Modell vorliegen. Mit dem Test wird also ermittelt, ob das Auslassen einer Einflußvariable aus dem Modell, während alle anderen Einflußgrößen vorhanden sind, zu einer deutlichen Verschlechterung der Vorhersagefunktionen des Modells führen würde. Diese Werte sollten nicht beim Auslassen mehrerer Variablen gleichzeitig für ein nachfolgendes Modell verwendet werden. Alternierende kleinste Quadrate optimieren zudem die Quantifikationen; dies impliziert, dass diese Tests auf konservative Weise interpretiert werden müssen.

Abbildung 9-19
Standardisierte Koeffizienten für transformierte Einflußgrößen

	Standardisierte Koeffizienten				
	Beta	Standardfehler	df	F	Sig.
Package design	-,748	,060	2	155,289	,000
Brand name	,045	,060	1	,578	,459
Price	,371	,059	1	39,312	,000
Good Housekeeping seal	-,350	,059	1	35,299	,000
Money-back guarantee	-,159	,059	1	7,175	,017

Abhängige Variable: Preference

Der größte Koeffizient liegt für *Package design* vor. Eine Erhöhung von *Package design* um eine Standardabweichung führt zu einem Abfall der vorhergesagten Einstufung der Bevorzugung um 0,748 Standardabweichungen. *Package design* wird jedoch nominal behandelt, sodass ein Anstieg der Quantifikationen nicht unbedingt mit einem Anstieg bei den ursprünglichen Kategoriecodes einhergeht.

Standardisierte Koeffizienten werden häufig fehlinterpretiert; es wird meist angenommen, dass diese Koeffizienten die Wichtigkeit der einzelnen Einflußgrößen widerspiegeln. Regressionskoeffizienten können jedoch die Auswirkungen einer Einflußgröße auf die Beziehungen zwischen den Einflußgrößen nicht vollständig beschreiben. Um die Auswirkungen von Einflußgrößen eingehend untersuchen zu können, müssen alternative Statistiken in Verbindung mit den standardisierten Koeffizienten verwendet werden.

Korrelationen und Wichtigkeit

Um den Beitrag der Einflußgrößen zur Regression korrekt zu interpretieren, reicht es nicht, wenn Sie lediglich die Regressionskoeffizienten betrachten. Darüber hinaus sollten auch die Korrelationen, die partiellen Korrelationen und die Teilkorrelationen berücksichtigt werden. Die nachstehende Tabelle zeigt diese korrelationalen Messwerte für die einzelnen Variablen.

Die Korrelation nullter Ordnung ist die Korrelation zwischen der transformierten Einflußgröße und der transformierten Antwort. Bei diesen Daten besteht die größte Korrelation bei *Package design*. Wenn Sie jedoch einen Teil der Variation bei der Einflußgröße oder der Antwort erklären können, erhalten Sie einen besseren Einblick, inwieweit die Einflußgröße ihren Zweck erfüllt.

Abbildung 9-20
Korrelationen nullter Ordnung, Teilkorrelationen und partielle Korrelationen (transformierte Variablen)

	Korrelationen				Toleranz	
	Nullte Ordnung	Partiell	Teil	Wichtigkeit	Nach Transformation	Vor Transformation
Package design	-,816	-,955	-,733	,644	,959	,942
Brand name	,206	,193	,045	,010	,971	,961
Price	,440	,851	,369	,172	,989	,982
Good Housekeeping seal	-,370	-,838	-,349	,137	,996	,991
Money-back guarantee	-,223	-,569	-,158	,037	,987	,993

Abhängige Variable: Preference

Andere Variablen im Modell können die Wichtigkeit einer gegebenen Einflußgröße bei der Vorhersage der Antwort durcheinanderbringen. Der partielle Korrelationskoeffizient beseitigt die linearen Auswirkungen anderer Einflußgrößen sowohl aus der Einflußgröße als auch aus der Antwort. Diese Messzahl entspricht der Korrelation zwischen den Residuen aus der Regression der Einflußgrößen über die anderen Einflußgrößen und den Residuen aus der Regression der Antwort über die anderen Einflußgrößen. Die quadrierte partielle Korrelation entspricht dem Anteil an der Varianz, der relativ zur residuellen Varianz der Antwort erklärt werden kann, die wiederum nach dem Beseitigen der Auswirkungen auf die anderen Variablen verbleibt. Beispiel: *Package design* weist eine partielle Korrelation von –0,955 auf. Werden die Auswirkungen der anderen Variablen entfernt, erklärt *Package design* $(–0,955)^2 = 0,91 = 91$ % der Variation in den Einstufungen der Bevorzugung. Auch *Price* und *Good Housekeeping seal* erklären einen Großteil der Varianz, wenn die Auswirkungen der anderen Variablen beseitigt werden.

Sie können die Auswirkungen anderer Variablen nicht nur aus der Antwort und aus einer Einflußgröße entfernen, sondern alternativ auch nur aus der Einflußgröße. Die Korrelation zwischen der Antwort und den Residuen aus der Regression einer Einflußgröße auf die anderen Einflußgrößen wird als Teilkorrelation bezeichnet. Wird dieser Wert quadriert, entsteht eine Messzahl für den Anteil der Varianz, der relativ zur Gesamtvarianz der Antwort erklärt wird. Wenn Sie die Auswirkungen von *Brand name*, *Good Housekeeping seal*, *Money back guarantee* und *Price* aus *Package design* entfernen, erklärt der verbleibende Teil von *Package design* $(–0,733)^2 = 0,54 = 54$ % der Variation bei den Einstufungen der Bevorzugung.

Wichtigkeit

Neben den Regressionskoeffizienten und den Korrelationen hilft das Maß für die relative Wichtigkeit nach Pratt (Pratt, 1987) bei der Interpretation des Beitrags der Einflußgrößen zur Regression. Hohe einzelne Wichtigkeitswerte relativ zu anderen Wichtigkeitswerten bezeichnen Einflußgrößen, die für die Regression von entscheidender Bedeutung sind. Darüber hinaus weist eine geringe Wichtigkeit für eine Variable, deren Koeffizient etwa gleich groß ist wie die wichtigen Einflußgrößen, auf die Präsenz von Suppressorvariablen hin.

Im Gegensatz zu den Regressionskoeffizienten definiert diese Messzahl die Wichtigkeit der Einflußgrößen additiv. Die Wichtigkeit einer Gruppe von Einflußgrößen entspricht also der Summe der einzelnen Werte für die Wichtigkeit dieser Einflußgrößen. Pratts Maß ist gleich dem Produkt des Regressionskoeffizienten und der Korrelation nullter Ordnung für eine Einflußgröße. Diese Produkte werden zu R^2 addiert und anschließend durch R^2 dividiert; dies ergibt die Summe 1. Die Gruppe der Einflußgrößen *Package design* und *Brand name* besitzt beispielsweise eine Wichtigkeit von 0,654. Die höchste Wichtigkeit liegt bei *Package design* vor; *Package design*,

Price und *Good Housekeeping seal* tragen 95 % der Wichtigkeit in dieser Kombination von Einflußgrößen bei.

Multi-Kollinearität

Hohe Korrelationen zwischen Einflußgrößen führen zu einer drastisch niedrigeren Stabilität des Regressionsmodells. Die korrelierten Einflußgrößen resultieren in instabilen Parameterschätzungen. Die Toleranz gibt wieder, inwieweit ein linearer Zusammenhang zwischen den unabhängigen Variablen besteht. Diese Messzahl ist der Anteil an der Varianz einer Variablen, der durch die anderen unabhängigen Variablen in der Gleichung nicht erklärt wird. Falls die anderen Einflußgrößen einen großen Teil der Varianz einer Einflußgröße erklären können, wird diese Einflußgröße im Modell nicht benötigt. Ein Toleranzwert nahe 1 bedeutet, dass die Variable nicht ohne weiteres aus den anderen Einflußgrößen vorhergesagt werden kann. Eine Variable mit sehr geringer Toleranz trägt dagegen wenig zum Informationsgehalt eines Modells bei und kann zu Problemen bei der Berechnung führen. Große negative Werte beim Maß für die Wichtigkeit nach Pratt weisen zudem auf Multi-Kollinearität hin.

Alle Toleranzmesswerte sind sehr hoch. Keine der Einflußgrößen kann ohne weiteres aus den anderen Einflußgrößen vorhergesagt werden und es besteht keine Multi-Kollinearität.

Transformationsdiagramme

Das Diagramm mit den ursprünglichen Werten und den zugehörigen Quantifikationen kann Trends aufdecken, die in einer Quantifikationsliste nicht ohne weiteres ersichtlich sind. Diese Diagramme werden in der Regel als Transformationsdiagramme bezeichnet. Kategorien mit ähnlichen Quantifikationen sollten besonders beachtet werden. Diese Kategorien wirken sich auf dieselbe Weise auf die vorhergesagte Antwort aus. Das grundlegende Erscheinungsbild des Diagramms wird allerdings durch den Transformationstyp bestimmt.

Variablen, die als numerisch behandelt werden, führen zu einer linearen Beziehung zwischen den Quantifikationen und den ursprünglichen Kategorien. Im Transformationsdiagramm wird dies als gerade Linie dargestellt. Die Reihenfolge der ursprünglichen Kategorien und die Differenzen zwischen diesen Kategorien werden in den Quantifikationen beibehalten.

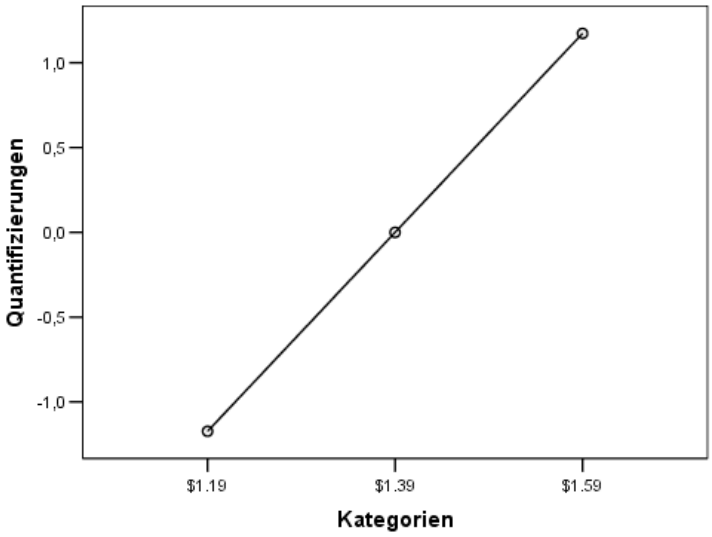
Die Reihenfolge der Quantifikationen für Variablen, die als ordinal behandelt werden, entspricht der Reihenfolge der ursprünglichen Kategorien. Die Differenzen zwischen den Kategorien werden allerdings nicht beibehalten. So entsteht ein nicht abnehmendes Transformationsdiagramm, das jedoch nicht unbedingt eine gerade Linie zeigen muss. Wenn aufeinanderfolgende Kategorien ähnlichen Quantifikationen entsprechen, ist die Unterscheidung der Kategorien ggf. nicht notwendig; die Kategorien könnten in diesem Fall kombiniert werden. Diese Kategorien bilden ein Plateau im Transformationsdiagramm. Dieses Muster kann jedoch auch entstehen, wenn eine ordinale Struktur über eine Variable, die als nominal behandelt werden sollte, überlagert wird. Zeigt eine nachfolgende nominale Behandlung der Variablen dasselbe Muster, ist die Kombination der Kategorien gerechtfertigt. Wenn die Quantifikationen für eine als ordinal behandelte Variable entlang einer geraden Linie liegen, dürfte darüber hinaus eine numerische Transformierung angebrachter sein.

Bei Variablen, die als nominal behandelt werden, ist die Reihenfolge der Kategorien an der horizontalen Achse mit der Reihenfolge der Codes identisch, die die Kategorien repräsentieren. Interpretationen der Kategorienreihenfolge oder der Distanz zwischen den Kategorien sind

gegenstandslos. Das Diagramm kann eine beliebige nichtlineare oder lineare Form annehmen. Besteht ein steigender Trend, sollte die ordinale Behandlung verfolgt werden. Falls das nominale Transformationsdiagramm einen linearen Trend zeigt, ist ggf. eine numerische Transformation angebrachter.

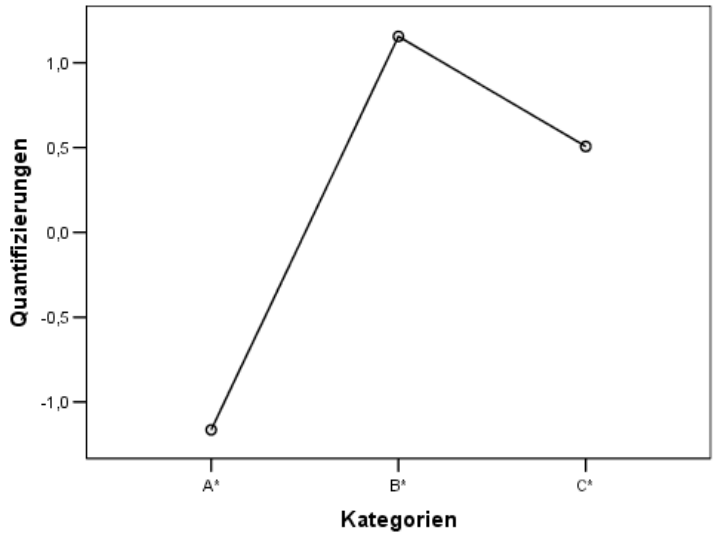
Die nachstehende Abbildung zeigt das Transformationsdiagramm für die Variable *Price*, die als numerisch behandelt wurde. Beachten Sie, dass die Reihenfolge der Kategorien entlang der geraden Linie mit der Reihenfolge der ursprünglichen Kategorien übereinstimmt. Darüber hinaus ist die Differenz zwischen den Quantifikationen für *$1.19* und *$1.39* (–1,173 und 0) mit der Differenz zwischen den Quantifikationen für *$1.39* und *$1.59* (0 und 1,173) identisch. Die Tatsache, dass die Kategorien 1 und 3 dieselbe Distanz zur Kategorie 2 besitzen, bleibt in den Quantifikationen erhalten.

Abbildung 9-21
Transformationsdiagramm für Preise (numerisch)



Die nominale Transformation von *Package design* führt zum nachstehenden Transformationsdiagramm. Beachten Sie die deutliche nichtlineare Form, in der die größte Quantifikation der zweiten Kategorie liegt. Im Hinblick auf die Regression senkt die zweite Kategorie die vorhergesagte Einstufung der Bevorzugung; die erste und dritte Kategorie zeigen dagegen den umgekehrten Effekt.

Abbildung 9-22
Transformationsdiagramm für "Package design" (nominal)

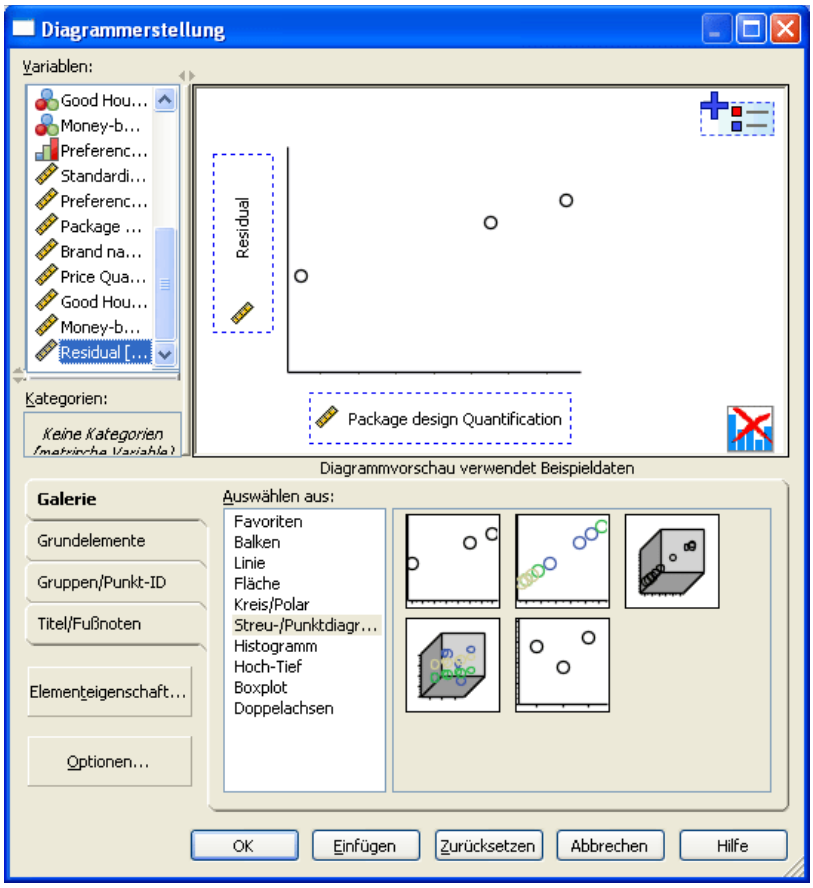


Analyse der Residuen

Mit den transformierten Daten und Residuen, die in der Arbeitsdatei gespeichert wurden, können Sie ein Streudiagramm der vorhergesagten Werte über die transformierten Werte von *Package design* erstellen.

Öffnen Sie hierzu erneut die Diagrammerstellung und klicken Sie auf Zurücksetzen. Die bisherige Auswahl wird gelöscht und die Standardoptionen werden wiederhergestellt.

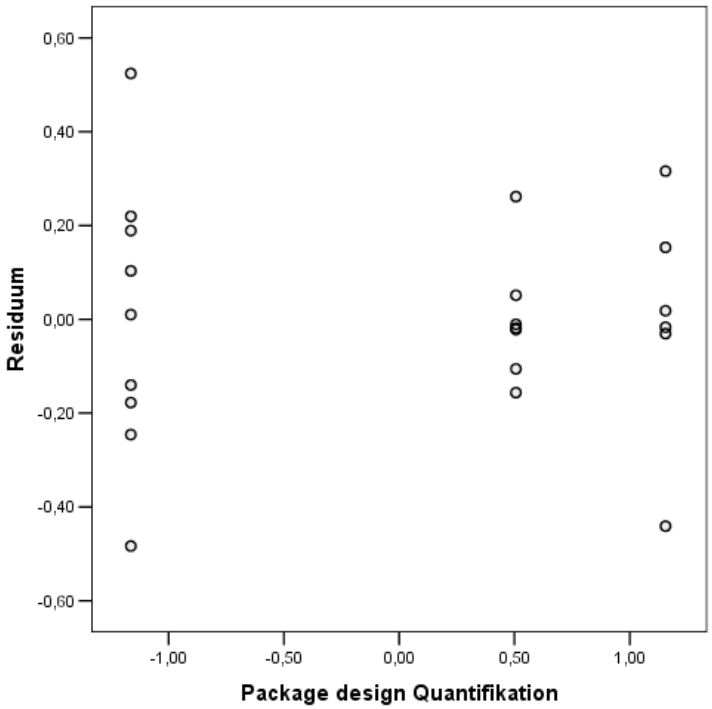
Abbildung 9-23
Diagrammerstellung



- Wählen Sie die Galerie Streu-/Punktdiagramm aus und wählen Sie die Option “Einfaches Streudiagramm”.
- Wählen Sie *Residual* als Variable für die *y*-Achse aus.
- Wählen Sie *Package design Quantification* als Variable für die *x*-Achse aus.
- Klicken Sie auf OK.

Das Streudiagramm zeigt die standardisierten Residuen anhand der optimalen Werte für *Package design*. Alle Residuen liegen innerhalb von zwei Standardabweichungen von 0. Die U-Form im Streudiagramm aus der linearen Regression nach dem Standardverfahren wird durch eine zufällige Streuung von Punkten ersetzt. Die Vorhersagekraft wird durch die optimale Quantifikation der Kategorien erreicht.

Abbildung 9-24
Residuen bei der kategorialen Regression



Beispiel: Ozondaten

In diesem Beispiel werden die Auswahl und die Auswirkungen von Transformationen mit optimaler Skalierung anhand eines größeren Daten-Sets verdeutlicht. Die Daten umfassen 330 Beobachtungen für sechs meteorologische Variablen, die unter anderem bereits von Breiman und Friedman (Breiman als auch Friedman, 1985) sowie von Hastie und Tibshirani (Hastie als auch Tibshirani, 1990) analysiert wurden. In der nachstehenden Tabelle werden die ursprünglichen Variablen beschrieben. Die kategoriale Regression soll die Ozonkonzentration aus den verbleibenden Variablen vorhersagen. Bei früheren Untersuchungen fanden Wissenschaftler einige Nichtlinearitäten unter diesen Variablen, die die Standardverfahren bei der Regression behindern.

Tabelle 9-2
Ursprüngliche Variablen

Variable	Beschreibung
ozon	daily ozone level; kategorisiert in eine von 38 Kategorien
ibh	inversion base height
dpg	pressure gradient (mm Hg)
vis	visibility (miles)
temp	temperature (degrees F)
doy	day of the year

Dieses Daten-Set finden Sie in *ozone.sav*. Für weitere Informationen siehe Beispieldateien in Anhang A auf S. 326.

Diskretisierung von Variablen

Wenn eine Variable so viele Kategorien enthält, dass eine Interpretation nicht mehr praktikabel wäre, bearbeiten Sie die Kategorien im Dialogfeld “Discretization”.

Für die Variable *Day of the year* gilt ein Mindestwert von 3 und ein Höchstwert von 365. Wird diese Variable in einer kategorialen Regression eingesetzt, entspricht dies einer Variablen mit 365 Kategorien. Bei *Visibility (miles)* gilt ein Bereich von 0 bis 350. Um die Interpretation der Analysen zu erleichtern, diskretisieren Sie diese Variablen in gleiche Intervalle mit der Länge 10.

Die Variable *Inversion base height* erstreckt sich über einen Bereich von 111 bis 5000. Eine Variable mit so vielen Kategorien resultiert in äußerst komplexen Beziehungen. Wenn Sie diese Variable jedoch in gleiche Intervalle mit der Länge 100 diskretisieren, erhalten Sie etwa 50 Kategorien. Eine Variable mit “nur” 50 Kategorien vereinfacht die Interpretation gegenüber einer Variablen mit 5000 Kategorien ganz erheblich.

Für *Pressure gradient (mm Hg)* gilt ein Bereich von –69 bis 107. Bei der Prozedur werden alle Kategorien aus der Analyse ausgeschlossen, die mit einem negativen Wert kodiert sind. Wenn Sie diese Variable jedoch in gleiche Intervalle mit der Länge 10 diskretisieren, ergeben sich etwa 19 Kategorien.

Temperature (degrees F) erstreckt sich von 25 bis 93 auf der Fahrenheit-Temperaturskala. Sollen diese Daten wie Daten auf der Celsius-Temperaturskala analysiert werden, diskretisieren Sie diese Variablen in gleiche Intervalle mit der Länge 1,8.

Variablen können auf unterschiedliche Weise diskretisiert werden. Die hier angegebenen Werte sind rein subjektiv. Wenn Sie weniger Kategorien erzielen möchten, verwenden Sie längere Intervalle. *Day of the year* könnte beispielsweise in Monate oder in Jahreszeiten aufgegliedert werden.

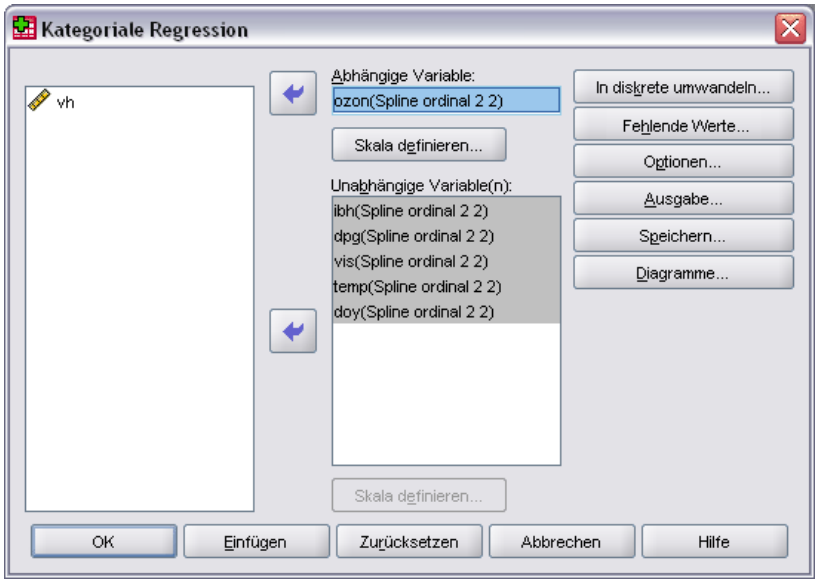
Auswahl des Transformationstyps

Jede Variable kann in einer von mehreren verschiedenen Ebenen analysiert werden. Das Ziel ist jedoch die Vorhersage der Antwort; darum sollten Sie die Antwort direkt mithilfe des numerischen Messniveaus der optimalen Skalierung skalieren. Die Reihenfolge und die Differenzen zwischen den Kategorien werden entsprechend in der transformierten Variablen beibehalten.

▶ Zum Ausführen einer kategorialen Regressionsanalyse wählen Sie die folgenden Befehle aus den Menüs aus:

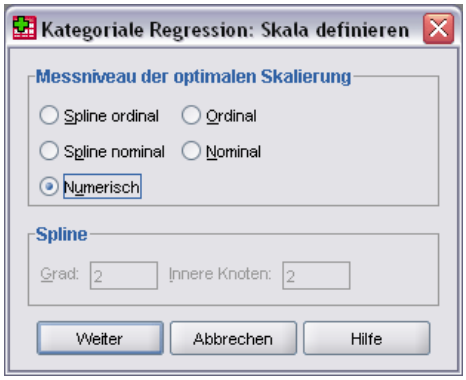
Analysieren
  Regression
    Optimale Skalierung...

Abbildung 9-25
Dialogfeld "Kategoriale Regression"



- Wählen Sie *Daily ozone level* als abhängige Variable aus.
- Wählen Sie *Inversion base height* bis *Day of the year* als unabhängige Variablen aus.
- Wählen Sie *Daily ozone level* und klicken Sie auf Skala definieren.

Abbildung 9-26
Dialogfeld "Skala definieren"



- Wählen Sie Numerisch als optimales Messniveau für die Skalierung.
- Klicken Sie auf Weiter.
- Wählen Sie *Inversion base height* bis *Day of the year* aus und klicken Sie im Dialogfeld "Kategoriale Regression" auf Skala definieren.

Abbildung 9-27
Dialogfeld "Skala definieren"

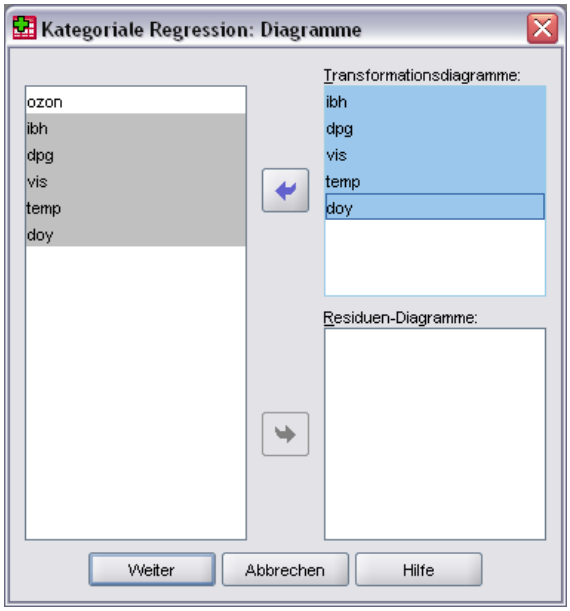
- Wählen Sie Nominal als optimales Messniveau für die Skalierung.
- Klicken Sie auf Weiter.
- Klicken Sie im Dialogfeld "Kategoriale Regression" auf In diskrete umwandeln.

Abbildung 9-28
Dialogfeld "Diskretisierung"

- Wählen Sie *ibh* aus.
- Wählen Sie Gleiche Intervalle und geben Sie 100 als Intervall ein.
- Klicken Sie auf Ändern.
- Wählen Sie *dpg*, *vis* und *doy* aus.

- Geben Sie 10 als Länge für das Intervall ein.
- Klicken Sie auf Ändern.
- Wählen Sie *temp* aus.
- Geben Sie 1,8 als Länge für das Intervall ein.
- Klicken Sie auf Ändern.
- Klicken Sie auf Weiter.
- Klicken Sie im Dialogfeld "Kategoriale Regression" auf Diagramme.

Abbildung 9-29
Dialogfeld "Diagramme"



- Wählen Sie Transformationsdiagramme für *Inversion base height* bis *Day of the year* aus.
- Klicken Sie auf Weiter.
- Klicken Sie im Dialogfeld "Kategoriale Regression" auf OK.

Abbildung 9-30
Modellzusammenfassung

Multiples R	R-Quadrat	Korrigiertes R-Quadrat
,941	,886	,793

Abhängige Variable: Daily ozone level
Einflußvariablen: Inversion base height Pressure gradient (mm Hg) Visibility (miles) Temperature (degrees F) Day of the year

Werden alle Einflußgrößen als nominal behandelt, ergibt sich ein Wert von 0,886 für R^2. Ein so großer Wert für die berücksichtigte Varianz ist nicht weiter überraschend, weil bei der nominalen Behandlung keine Einschränkungen für die Quantifikationen auftreten. Die Interpretation der Ergebnisse kann sich allerdings recht schwierig gestalten.

Abbildung 9-31
Regressionskoeffizienten (alle Einflußgrößen nominal)

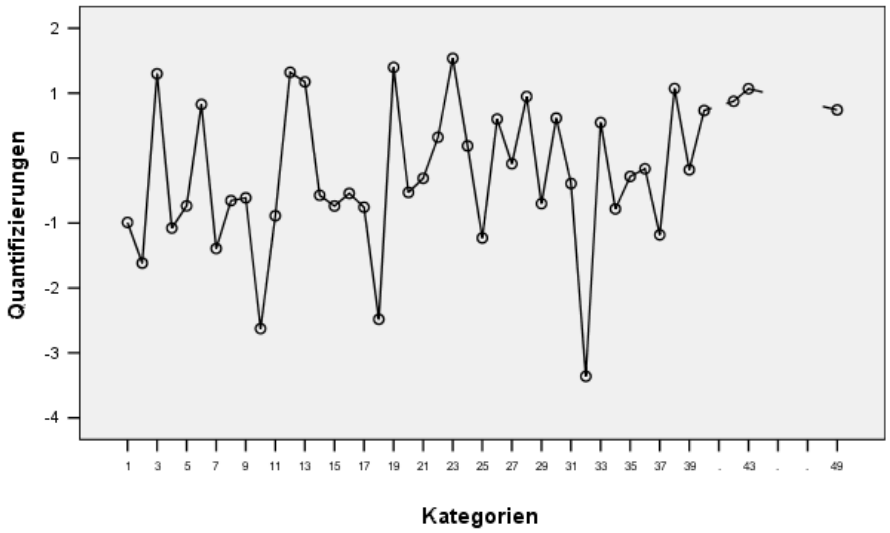
	Standardisierte Koeffizienten				
	Beta	Standard fehler	df	F	Sig.
Inversion base height	-,309	,026	42	144,187	,000
Pressure gradient (mm Hg)	,307	,028	16	123,227	,000
Visibility (miles)	-,216	,026	17	69,528	,000
Temperature (degrees F)	,588	,027	36	468,542	,000
Day of the year	-,408	,029	36	203,250	,000

Abhängige Variable: Daily ozone level

Diese Tabelle ziegt die standardisierten Regressionskoeffizienten für die Einflußgrößen. Ein häufiger Fehler beim Interpretieren dieser Werte ist es, den Schwerpunkt auf die Koeffizienten zu legen und dabei die Quantifikationen zu vernachlässigen. Sie können nicht voraussetzen, dass der große positive Wert für den Koeffizienten *Temperature* impliziert, dass der vorhergesagte Wert für *Ozone* automastisch steigt, wenn die Temperatur ansteigt.

Umgekehrt läßt der negative Koeffizient für *Inversion base height* nicht darauf schließen, dass ein Anstieg bei *Inversion base height* zwangsläufig zu einem Absinken des vorhergesagten Werts für *Ozone* führt. Die gesamte Interpretation muss relativ zu den transformierten Variablen vorgenommen werden. Wenn die Quantifikationen für *Temperature* steigen oder die Quantifikationen für *Inversion base height* sinken, steigt der vorhergesagte Wert für *Ozone*. Um die Auswirkungen der ursprünglichen Variablen untersuchen zu können, müssen die Kategorien in einen Zusammenhang mit den Quantifikationen gebracht werden.

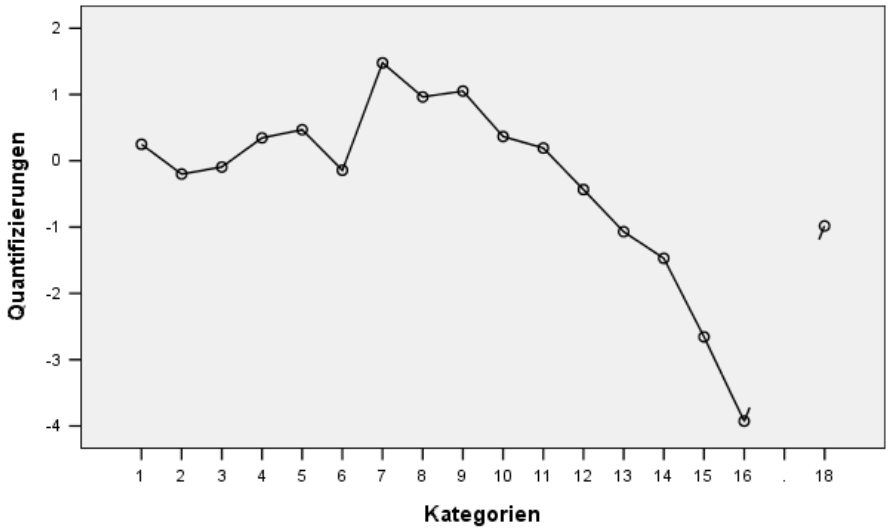
Abbildung 9-32
Transformationsdiagramm von "Inversion base height" (nominal)



Das Transformationsdiagramm von *Inversion base height* zeigt kein erkennbares Muster. Der gezackte Verlauf des Diagramms weist darauf hin, dass der Übergang von niedrigen zu hohen Kategorien in Fluktuationen bei den Quantifikationen in beide Richtungen führt. Eine Beschreibung der Auswirkungen dieser Variablen muss sich daher auf die einzelnen Kategorien

konzentrieren. Wenn Sie ordinale oder lineare Einschränkungen für die Quantifikationen dieser Variablen festlegen, geht die Anpassungsgüte unter Umständen deutlich zurück.

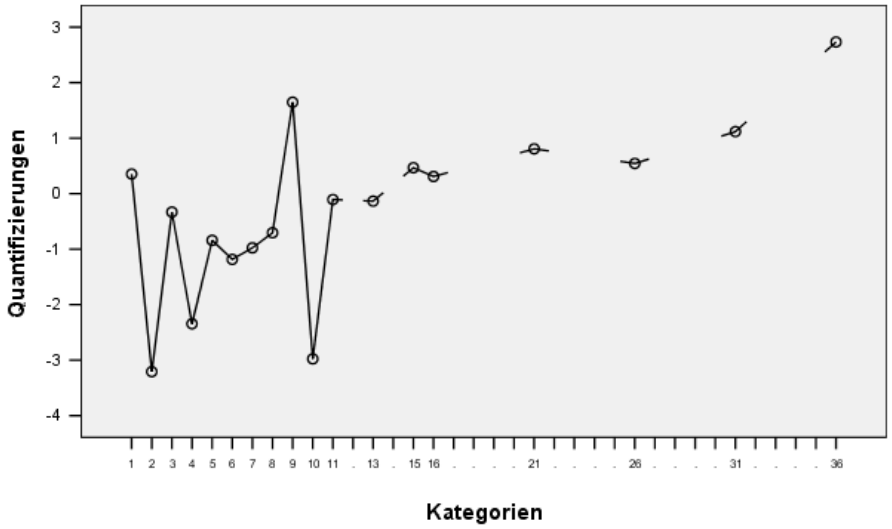
Abbildung 9-33
Transformationsdiagramm von "Pressure gradient" (nominal)



Diese Abbildung zeigt das Transformationsdiagramm von *Pressure gradient*. Die anfänglichen diskretisierten Kategorien (*1* bis *6*) erhalten nur kleine Quantifikationen und leisten daher nur einen minimalen Beitrag zur vorhergesagten Antwort. Die nächsten drei Kategorien erhalten etwas höhere, positive Werte. Dies führt zu einem mäßigen Anstieg beim vorhergesagten Ozonwert.

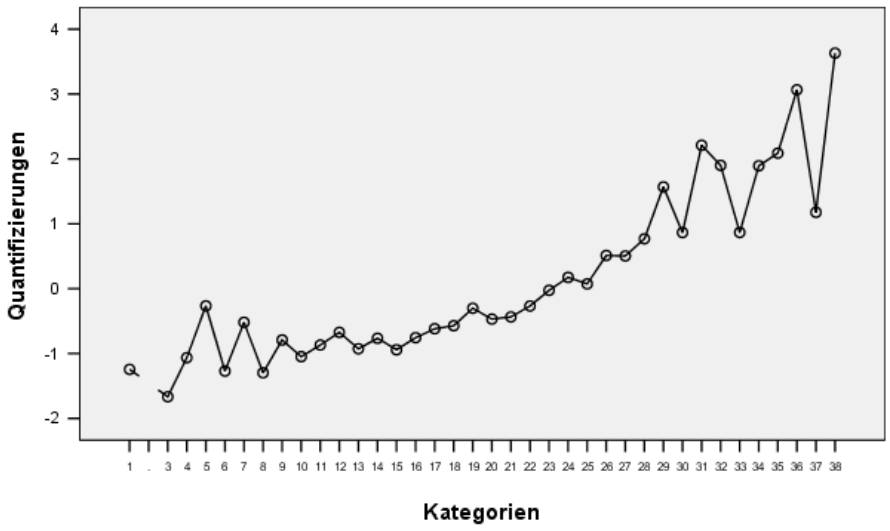
Die Quantifikationen sinken weiter bis zur Kategorie *16*, in der *Pressure gradient* den größten absenkenden Effekt auf den vorhergesagten Ozonwert hat. Die Linie fällt auch nach dieser Kategorie noch weiter ab; ein ordinales Messniveau für die Skalierung von *Pressure gradient* dürfte die Anpassungsgüte nicht merklich verringern, vereinfacht jedoch die Interpretation der Auswirkungen. Das Maß 0,4 für die Wichtigkeit und der Regressionskoeffizient für *Pressure gradient* weisen allerdings darauf hin, dass diese Variable bei der Regression nicht sonderlich nützlich ist.

Abbildung 9-34
Transformationsdiagramm von "Visibility" (nominal)



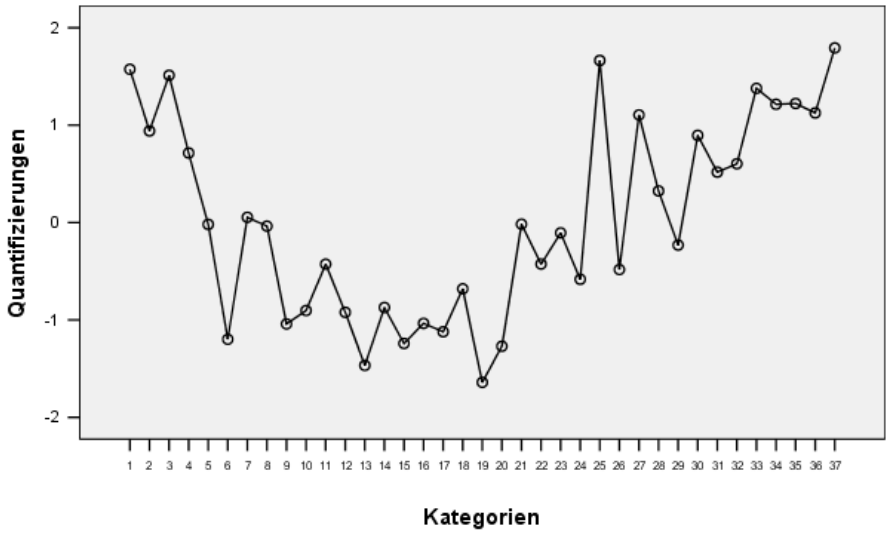
Das Transformationsdiagramm von *Visibility* zeigt wie das Transformationsdiagramm für *Inversion base height* kein erkennbares Muster. Wenn Sie ordinale oder lineare Einschränkungen für die Quantifikationen dieser Variablen festlegen, geht die Anpassungsgüte unter Umständen deutlich zurück.

Abbildung 9-35
Transformationsdiagramm von "Temperature" (nominal)



Das Transformationsdiagramm von *Temperature* zeigt ein anderes Muster. In höheren Kategorien steigen auch die Quantifikationen. Das bedeutet: Wenn *Temperature* ansteigt, steigen auch die vorhergesagten Ozonwerte. Aus diesem Muster geht hervor, dass eine Skalierung von *Temperature* auf ordinalem Messniveau empfehlenswert wäre.

Abbildung 9-36
Transformationsdiagramm von "Day of the year" (nominal)

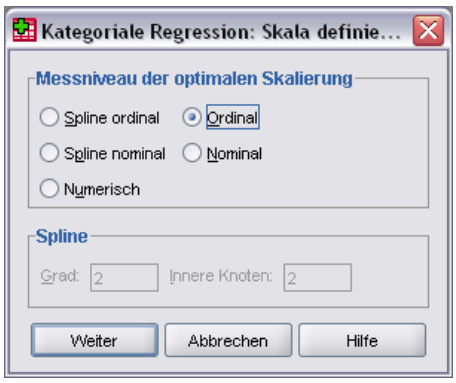


Diese Abbildung zeigt das Transformationsdiagramm von *Day of the year*. Die Quantifikationen steigen bis Kategorie *19* an und sinken dann wieder ab. So ergibt sich eine U-Form. Unter Berücksichtigung des Vorzeichens beim Regressionskoeffizienten für *Day of the year* erhalten die ursprünglichen Kategorien (*1* bis *5*) jeweils eine Quantifikation, die sich senkend auf den vorhergesagten Ozonwert auswirkt. Ab Kategorie *6* entstehen steigernde Auswirkungen auf die Quantifikationen, bis etwa bei Kategorie *19* der Höchstwert erreicht wird.

Ab Kategorie *19* führen die Quantifikationen zu einer Senkung des vorhergesagten Ozonwerts. Die Linie ist zwar recht gezackt, läßt jedoch die zugrunde liegende Form immer noch erkennen. Die Transformationsdiagramme zeigen daher, dass *Temperature* auf ordinaler Ebene skaliert werden sollte; für alle anderen Einflußgrößen sollte statt dessen die nominale Skalierung verwendet werden.

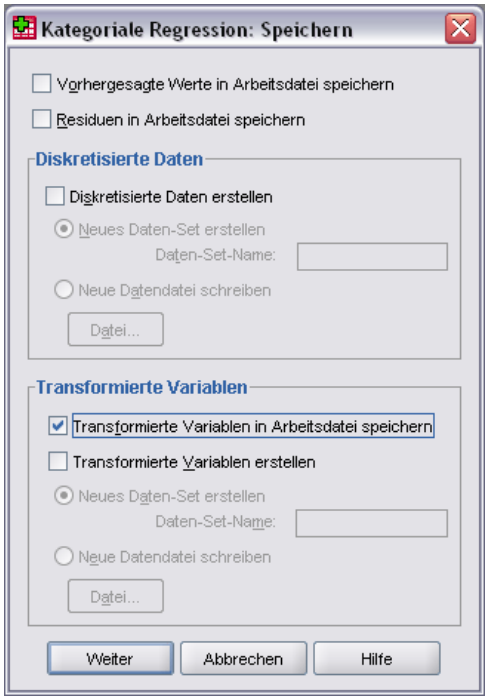
Zum Neuberechnen der Regression (mit Skalierung von *Temperature* auf ordinaler Ebene) öffnen Sie erneut das Dialogfeld "Kategoriale Regression".

Abbildung 9-37
Dialogfeld "Skala definieren"



- Wählen Sie *Temperature* aus und klicken Sie auf Skala definieren.
- Wählen Sie als Messniveau der optimalen Skalierung die Option Ordinal aus.
- Klicken Sie auf Weiter.
- Klicken Sie im Dialogfeld "Kategoriale Regression" auf Speichern.

Abbildung 9-38
Dialogfeld "Speichern"



- Wählen Sie Im Gruppenfeld "Transformierte Variablen" die Option Transformierte Variablen in Arbeitsdatei speichern aus.
- Klicken Sie auf Weiter.
- Klicken Sie im Dialogfeld "Kategoriale Regression" auf OK.

Abbildung 9-39
Modellzusammenfassung für die Regression mit "Temperature" (ordinal)

Multiples R	R-Quadrat	Korrigiertes R-Quadrat
,935	,875	,791

Abhängige Variable: Daily ozone level
Einflußvariablen: Inversion base height Pressure gradient (mm Hg) Visibility (miles) Temperature (degrees F) Day of the year

Dieses Modell ergibt einen Wert von 0,875 für R^2. Die berücksichtigte Varianz fällt daher geringfügig ab, wenn die Quantifikationen für *Temperature* in eine Reihenfolge gebracht werden.

Abbildung 9-40
Regressionskoeffizienten mit "Temperature" (ordinal)

	Standardisierte Koeffizienten				
	Beta	Standardfehler	df	F	Sig.
Inversion base height	-,300	,026	42	132,628	,000
Pressure gradient (mm Hg)	,294	,028	16	111,757	,000
Visibility (miles)	-,221	,026	17	72,360	,000
Temperature (degrees F)	,619	,027	20	517,262	,000
Day of the year	-,372	,028	36	173,158	,000

Abhängige Variable: Daily ozone level

Diese Tabelle zeigt die Koeffizienten für das Modell, in dem *Temperature* als ordinal skaliert wurde. Der Vergleich der Koeffizienten mit den Koeffizienten im Modell, bei denen *Temperature* als nominal skaliert ist, zeigt keine weitreichenden Veränderungen.

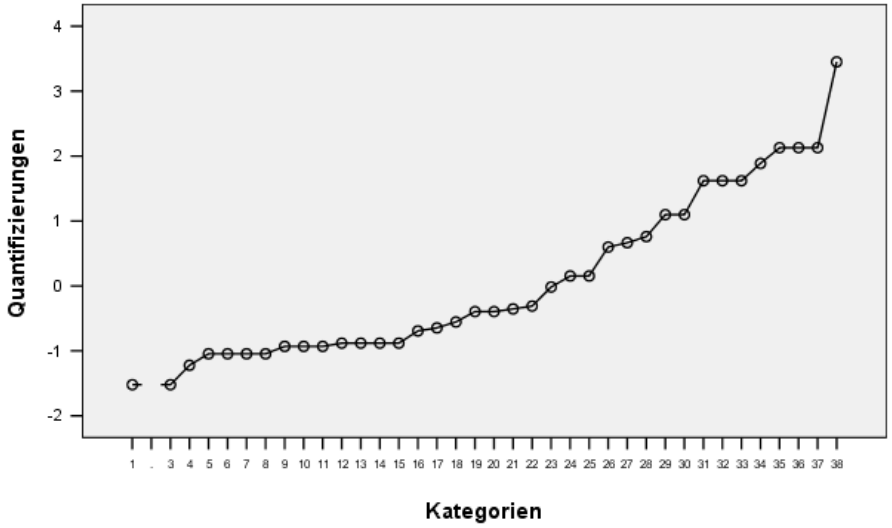
Abbildung 9-41
Korrelationen, Wichtigkeit und Toleranz

	Korrelationen				Toleranz	
	Nullte Ordnung	Partiell	Teil	Wichtigkeit	Nach Transformation	Vor Transformation
Inversion base height	-,435	-,633	-,290	,150	,931	,596
Pressure gradient (mm Hg)	,135	,601	,266	,045	,819	,859
Visibility (miles)	-,352	-,517	-,214	,089	,939	,752
Temperature (degrees F)	,806	,850	,573	,571	,855	,580
Day of the year	-,340	-,683	-,331	,145	,791	,801

Abhängige Variable: Daily ozone level

Die Maße für die Wichtigkeit lassen zudem erkennen, dass *Temperature* immer noch von größerer Bedeutung für die Regression ist als die anderen Variablen. Als Ergebnis des Messniveaus mit ordinaler Skalierung für *Temperature* und des positiven Regressionskoeffizienten können Sie nun jedoch voraussetzen, dass der vorhergesagte Ozonwert steigt, sobald *Temperature* ansteigt.

Abbildung 9-42
Transformationsdiagramm von "Temperature" (ordinal)



Das Transformationsdiagramm zeigt die ordinale Einschränkung der Quantifikationen für *Temperature*. Die gezackte Linie aus der nominalen Transformationen ist hier durch eine glatte, ansteigende Linie ersetzt. Außerdem sind keine langen Plateaus mehr vorhanden. Dies bedeutet, dass eine Zusammenfassung der Kategorien nicht erforderlich ist.

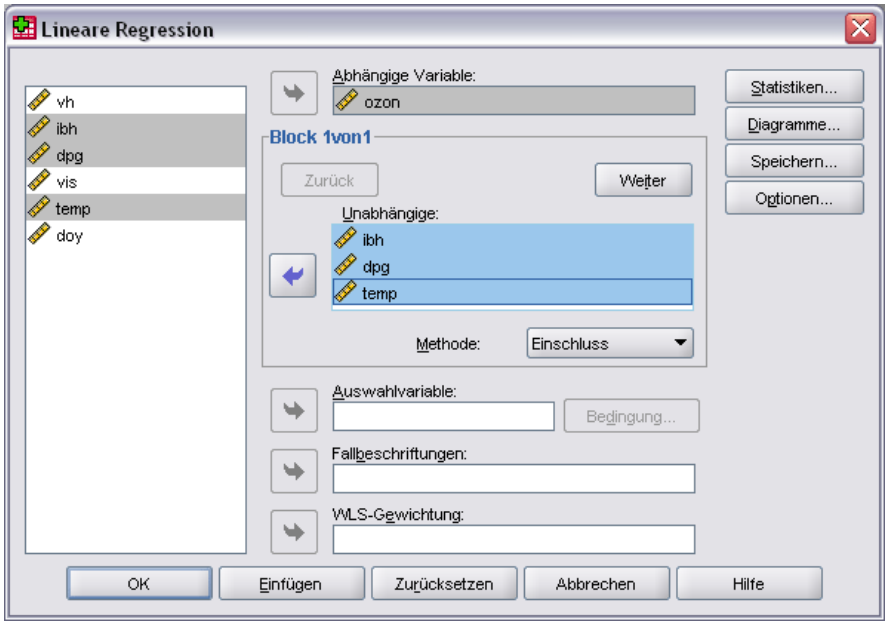
Optimalität der Quantifikationen

Die transformierten Variablen aus einer kategorialen Regression können bei einer linearen Regression im Standardverfahren genutzt werden; dabei entstehen identische Ergebnisse. Die Quantifikationen sind jedoch nur für das Modell optimal, mit dem sie erzeugt wurden. Die Verwendung einer Untergruppe der Einflußgrößen in einer linearen Regression ist nicht identisch mit einer Regression mit optimaler Skalierung für diese Untergruppe.

Die berechnete kategoriale Regression besitzt beispielsweise einen Wert von 0,875 für R^2. Sie haben die transformierten Variablen gespeichert. Um also die Anpassungsgüte einer linearen Regression nur mit *Temperature*, *Pressure gradient* und *Inversion base height* als Einflußgrößen zu bestimmen, wählen Sie aus den Menüs die folgenden Befehle aus:

Analysieren
  Regression
    Linear...

Abbildung 9-43
Dialogfeld "Lineare Regression"



- Wählen Sie *Daily ozone level Quantification* als abhängige Variable aus.
- Wählen Sie *Inversion base height Quantification*, *Pressure gradient (mm Hg) Quantification* und *Temperature (degrees F) Quantification* als unabhängige Variablen aus.
- Klicken Sie auf OK.

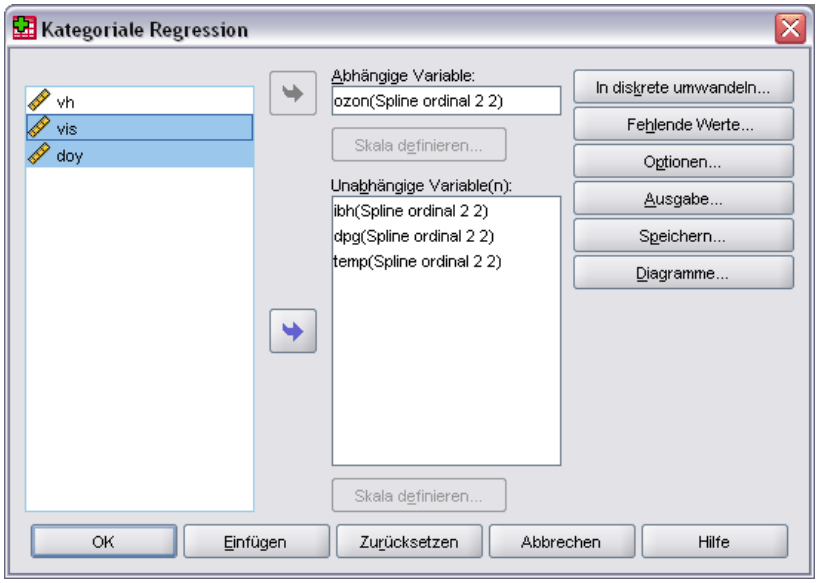
Abbildung 9-44
Modellzusammenfassung für die Regression mit einer Untergruppe von Einflußgrößen mit optimaler Skalierung

Modell	R	R-Quadrat	Korrigiertes R-Quadrat	Standard fehler des Schätzers
1	,856[a]	,733	,730	,51993

a. Einflußvariablen : (Konstante), Temperature (degrees F) Quantifikation, Pressure gradient (mm Hg) Quantifikation, Inversion base height Quantifikation

Unter Verwendung der Quantifikationen für die Antwort führen *Temperature*, *Pressure gradient* und *Inversion base height* in einer linearen Regression nach dem Standardverfahren zu einer Anpassungsgüte von 0,733. Soll dieser Wert mit der Anpassungsgüte einer kategorialen Regression verglichen werden, die sich ebenfalls nur auf diese drei Einflußgrößen erstreckt, öffnen Sie erneut das Dialogfeld "Kategoriale Regression".

Abbildung 9-45
Dialogfeld "Kategoriale Regression"



- Deaktivieren Sie die Einträge *Visibility (miles)* und *Day of the year* als unabhängige Variablen.
- Klicken Sie auf OK.

Abbildung 9-46
Modellzusammenfassung für die kategoriale Regression mit drei Einflußgrößen

Multiples R	R-Quadrat	Korrigiertes R-Quadrat
,893	,798	,740

Abhängige Variable: Daily ozone level
Einflußvariablen: Inversion base height Pressure gradient (mm Hg) Temperature (degrees F)

Die kategoriale Regressionsanalyse ergibt eine Anpassungsgüte von 0,798, also einen besseren Wert als bei der Güte 0.733. Dies belegt die Eigenschaft der Skalierung, dass die in der ursprünglichen Regression erzielten Quantifikationen nur dann optimal sind, wenn alle fünf Variablen im Modell berücksichtigt werden.

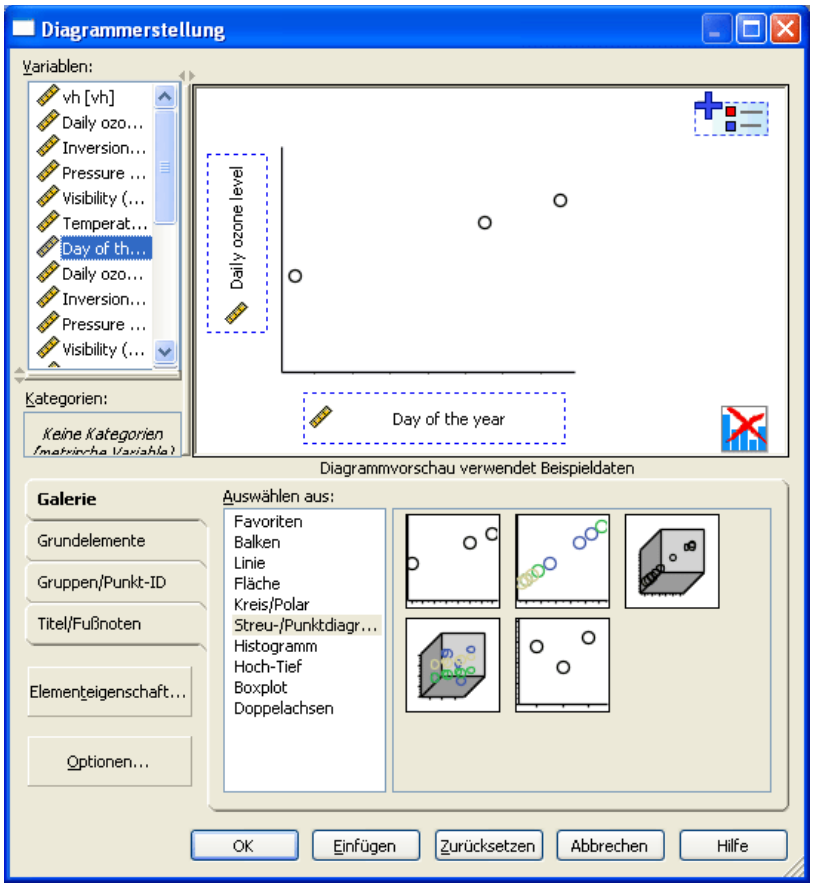
Auswirkungen von Transformationen

Beim Transformieren der Variablen wird eine nichtlineare Beziehung zwischen der ursprünglichen Antwort und der ursprünglichen Gruppe von Einflußgrößen in eine lineare Beziehung für die transformierten Variablen umgewandelt. Wenn jedoch mehrere Einflußgrößen vorliegen, werden paarweise Beziehungen durch die anderen Variablen im Modell durcheinandergebracht.

Soll sich die Analyse auf die Beziehung zwischen *Daily ozone level* und *Day of the year* konzentrieren, betrachten Sie zunächst ein Streudiagramm. Wählen Sie die folgenden Befehle aus den Menüs aus:

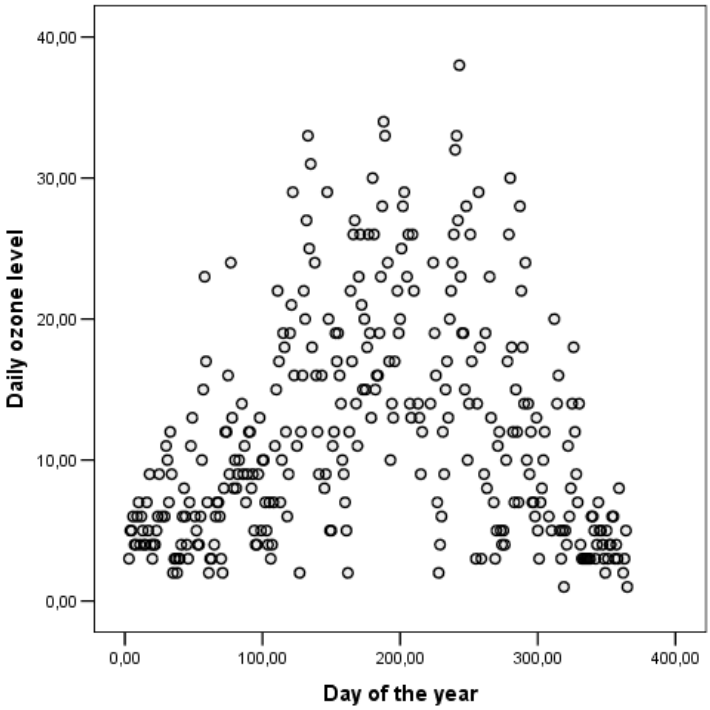
Grafiken
Diagrammerstellung...

Abbildung 9-47
Dialogfeld "Diagrammerstellung"



- Wählen Sie die Galerie Streu-/Punktdiagramm aus und wählen Sie die Option "Einfaches Streudiagramm".
- Wählen Sie *Daily ozone level* als *y*-Achsen-Variable aus und *Day of the year* als *x*-Achsen-Variable.
- Klicken Sie auf OK.

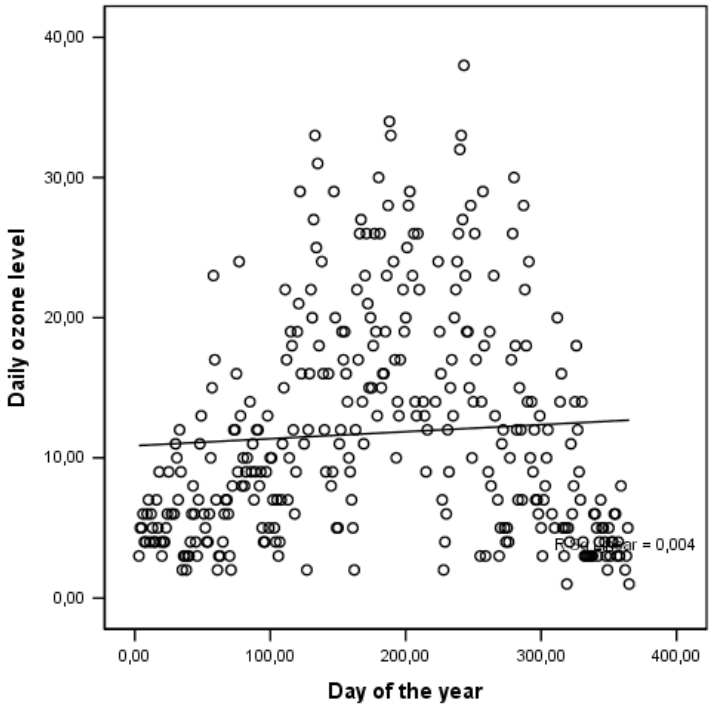
Abbildung 9-48
Streudiagramm von "Daily ozone level" und "Day of the year"



Diese Abbildung zeigt das Verhältnis zwischen *Daily ozone level* und *Day of the year*. Wenn *Day of the year* auf etwa 200 ansteigt, steigt auch *Daily ozone level*. Bei Werten größer 200 für *Day of the year* sinkt *Daily ozone level* jedoch ab. Diese invertierte U-Form weist auf eine quadratische Beziehung zwischen den beiden Variablen hin. Eine lineare Beziehung kann diese Beziehung nicht erfassen.

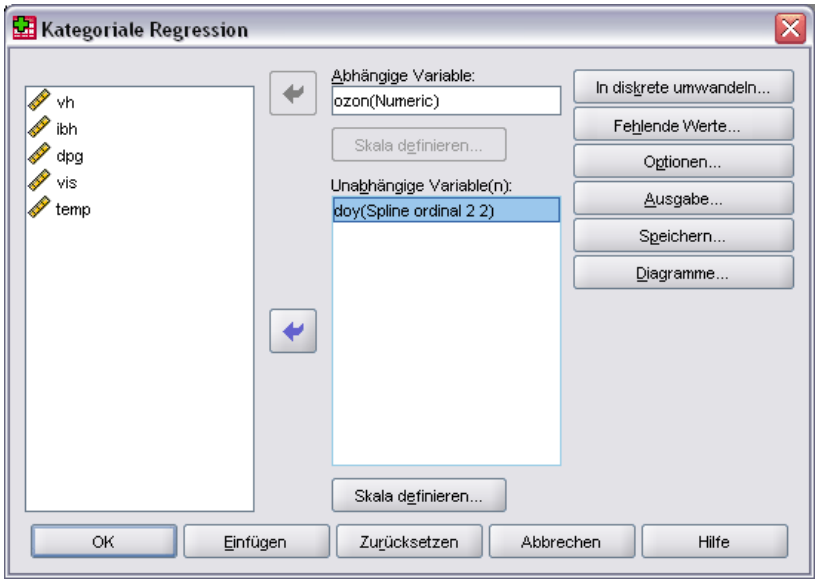
- Um die am besten geeignete Linie über die Punkte im Streudiagramm anzeigen zu lassen, aktivieren Sie die Grafik, indem Sie darauf doppelklicken.
- Wählen Sie im Diagramm-Editor einen Punkt aus.
- Klicken Sie auf das Werkzeug zum Einfügen einer Anpassungslinie beim Gesamtwert und schließen Sie den Diagramm-Editor.

Abbildung 9-49
Streudiagramm mit Anpassungslinie



Eine lineare Regression von *Daily ozone level* über *Day of the year* ergibt einen Wert von 0,004 für R^2. Diese Anpassungsgüte läßt erkennen, dass *Day of the year* keinen vorhersagbaren Wert für *Daily ozone level* aufweist. Beim vorliegenden Muster in der Abbildung ist dies nicht weiter überraschend. Wenn Sie jedoch die optimale Skalierung verwenden, können Sie die quadratische Beziehung linearisieren und die Antwort mithilfe der transformierten Variablen *Day of the year* vorhersagen.

Abbildung 9-50
Dialogfeld "Kategoriale Regression"



Soll eine kategoriale Regression von *Daily ozone level* über *Day of the year* erstellt werden, rufen Sie erneut das Dialogfeld "Kategoriale Regression" auf.

- Deaktivieren Sie *Inversion base height* bis *Temperature (degrees F)* als unabhängige Variablen.
- Wählen Sie *Day of the year* als unabhängige Variable aus.
- Klicken Sie auf Skala definieren.

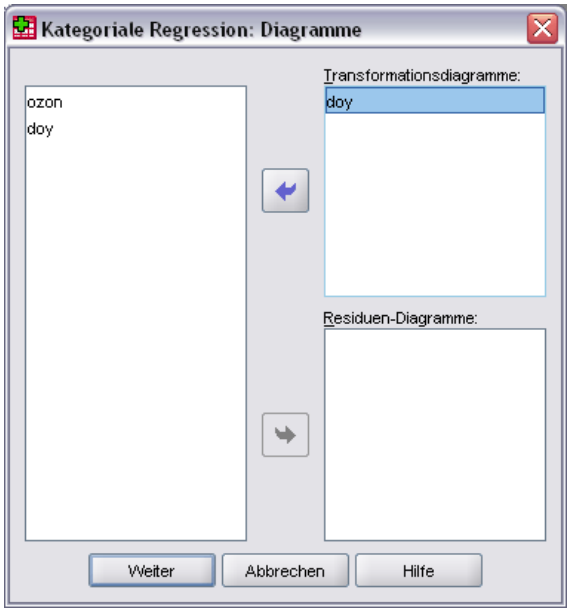
Abbildung 9-51
Dialogfeld "Skala definieren"

- Wählen Sie Nominal als optimales Messniveau für die Skalierung.
- Klicken Sie auf Weiter.
- Klicken Sie im Dialogfeld "Kategoriale Regression" auf In diskrete umwandeln.

Abbildung 9-52
Dialogfeld "Diskretisierung"

- Wählen Sie *doy* aus.
- Wählen Sie Gleiche Intervalle.
- Geben Sie 10 als Länge für das Intervall ein.
- Klicken Sie auf Ändern.
- Klicken Sie auf Weiter.
- Klicken Sie im Dialogfeld "Kategoriale Regression" auf Diagramme.

Abbildung 9-53
Dialogfeld "Diagramme"



- Wählen Sie *doy* für die Transformationsdiagramme aus.
- Klicken Sie auf Weiter.
- Klicken Sie im Dialogfeld "Kategoriale Regression" auf OK.

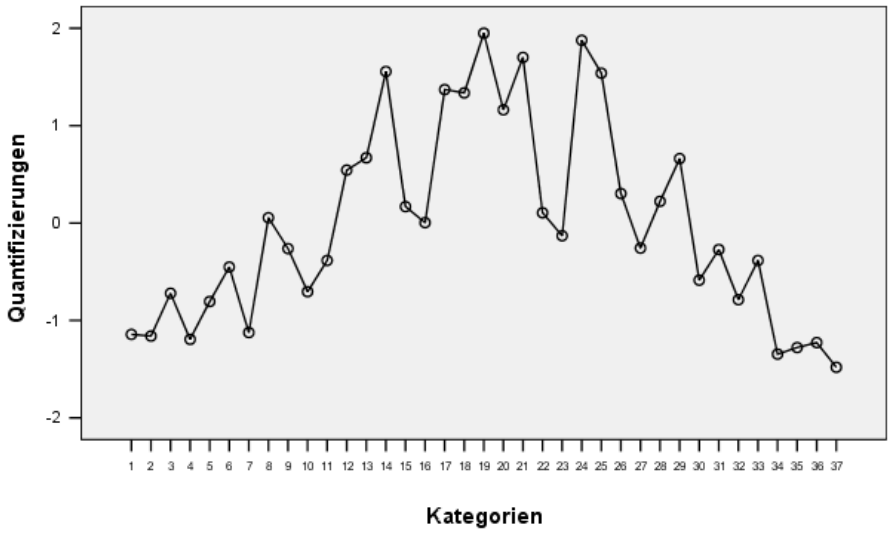
Abbildung 9-54
Modellzusammenfassung für die kategoriale Zusammenfassung von Daily ozone level auf Day of the year

Multiples R	R-Quadrat	Korrigiertes R-Quadrat
,741	,549	,494

Abhängige Variable: Daily ozone level
Einflußvariablen: Day of the year

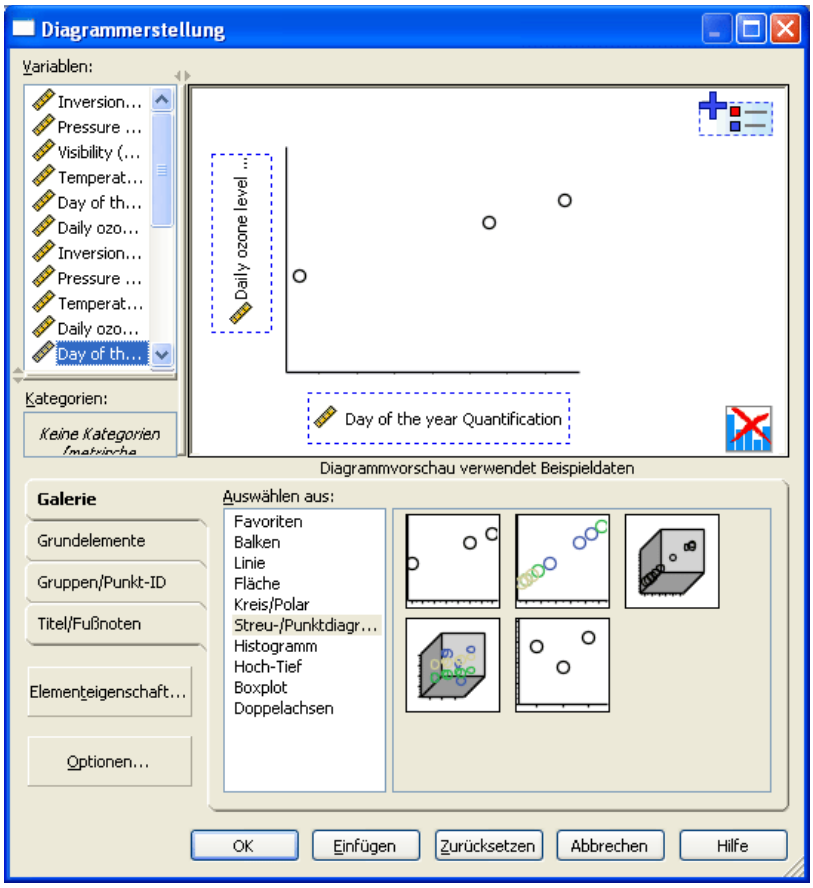
Bei der Regression mit optimaler Skalierung wird *Daily ozone level* als numerisch behandelt und *Day of the year* als nominal. Dies führt zu einem Wert von 0,549 für R^2. Obwohl nur 55 % der Variation bei *Daily ozone level* in der kategorialen Regression berücksichtigt sind, ist dies eine deutliche Verbesserung gegenüber der ursprünglichen Regression. Wenn Sie *Day of the year* transformieren, kann *Daily ozone level* vorhergesagt werden.

Abbildung 9-55
Transformationsdiagramm von "Day of the year" (nominal)



Diese Abbildung zeigt das Transformationsdiagramm von *Day of the year*. Die Extremwerte für *Day of the year* erhalten jeweils negative Quantifikationen, die zentralen Werte besitzen dagegen positive Quantifikationen. Mithilfe dieser Transformation zeigen die hohen und niedrigen Werte für *Day of the year* ähnliche Auswirkungen auf den vorhergesagten Wert für *Daily ozone level*.

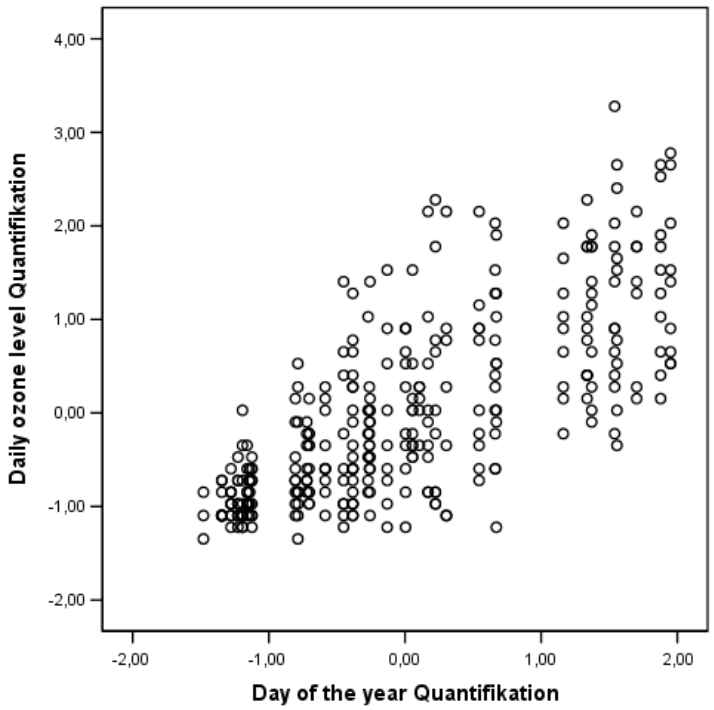
Abbildung 9-56
Diagrammerstellung



Zum Anzeigen eines Streudiagramms für die transformierten Variablen öffnen Sie erneut die Diagrammerstellung und klicken Sie auf Zurücksetzen. Die bisherige Auswahl wird gelöscht.

- Wählen Sie die Galerie Streu-/Punktdiagramm aus und wählen Sie die Option “Einfaches Streudiagramm”.
- Wählen Sie *Daily ozone level Quantification [TRA1_3]* als *y*-Achsen-Variable aus und *Day of the year Quantification [TRA2_3]* als *x*-Achsen-Variable.
- Klicken Sie auf OK.

Abbildung 9-57
Streudiagramm der transformierten Variablen



Diese Abbildung zeigt die Beziehungen zwischen den transformierten Variablen. Die umgekehrte U-Form wird durch einen steigenden Trend ersetzt. Die Regressionslinie besitzt eine positive Steigung; dies weist darauf hin, dass der vorhergesagte Wert für *Daily ozone level* steigt, wenn die transformierte Variable *Day of the year* ansteigt. Die Beziehung wird mithilfe der optimalen Skalierung linearisiert, sodass Interpretationen möglich sind, die ansonsten unbemerkt geblieben wären.

Empfohlene Literatur

In folgenden Texten finden Sie weitere Informationen zur kategorialen Regression:

Buja, A. 1990. Remarks on functional canonical variates, alternating least squares methods and ACE. *Annals of Statistics*, 18, 1032–1069.

Hastie, T., R. Tibshirani, als auch A. Buja. 1994. Flexible discriminant analysis. *Journal of the American Statistical Association*, 89, 1255–1270.

Hayashi, C. 1952. On the prediction of phenomena from qualitative data and the quantification of qualitative data from the mathematico-statistical point of view. *Annals of the Institute of Statitical Mathematics*, 2, 93–96.

Kruskal, J. B. 1965. Analysis of factorial experiments by estimating monotone transformations of the data. *Journal of the Royal Statistical Society Series B*, 27, 251–263.

Meulman, J. J. 2003. Prediction and classification in nonlinear data analysis: Something old, something new, something borrowed, something blue. *Psychometrika*, 4, 493–517.

Ramsay, J. O. 1989. Monotone regression splines in action. *Statistical Science*, 4, 425–441.

Van der Kooij, A. J., als auch J. J. Meulman. 1997. MURALS: Multiple regression and optimal scaling using alternating least squares. In: *Softstat '97,* F. Faulbaum, als auch W. Bandilla (Hgg.). Stuttgart: Gustav Fischer, 99–106.

Winsberg, S., als auch J. O. Ramsay. 1980. Monotonic transformations to additivity using splines. *Biometrika*, 67, 669–674.

Winsberg, S., als auch J. O. Ramsay. 1983. Monotone spline transformations for dimension reduction. *Psychometrika*, 48, 575–595.

Young, F. W., J. De Leeuw, als auch Y. Takane. 1976. Regression with qualitative and quantitative variables: An alternating least squares method with optimal scaling features. *Psychometrika*, 41, 505–528.

Kapitel
10

Kategoriale Hauptkomponentenanalyse

Die kategoriale Hauptkomponentenanalyse kann als Methode der Dimensionsreduktion betrachtet werden. Ein Variablen-Set wird analysiert, um größere Variationsdimensionen aufzudecken. Die ursprüngliche Datenmenge kann danach unter minimalem Informationsverlust durch eine neue, kleinere Datenmenge ersetzt werden. Die Methode zeigt Beziehungen zwischen Variablen, zwischen Fällen und zwischen Variablen und Fällen auf.

Das von der kategorialen Hauptkomponentenanalyse für die Quantifizierung der beobachteten Daten verwendete Kriterium besteht darin, dass die Objektwerte (Komponentenwerte) starke Korrelationen mit jeder der quantifizierten Variablen aufweisen sollten. Eine Lösung ist gut in dem Grad, in dem dieses Kriterium erfüllt wird.

Es werden zwei Beispiele der kategorialen Hauptkomponentenanalyse vorgestellt. Beim ersten wird eine eher kleine Datenmenge verwendet, anhand derer die Grundkonzepte und Interpretationen in Verbindung mit der Prozedur illustriert werden sollen. Im zweiten Beispiel wird eine praktische Anwendung untersucht.

Beispiel: Untersuchung der Interrelationen sozialer Systeme

In diesem Beispiel wird die Guttmansche (Guttman, 1968) Adaptation einer Tabelle von Bell (Bell, 1961) untersucht. Diese Daten werden auch von Lingoes (Lingoes, 1968) diskutiert.

Bell legte eine Tabelle zur Illustration möglicher sozialer Gruppen vor. Guttman verwendete einen Teil dieser Tabelle, bei der fünf Variablen, die Aspekte beschreiben, wie soziale Interaktion, das Gefühl der Gruppenzugehörigkeit, die physische Nähe der Mitglieder und die Formalität der Beziehung, mit sieben theoretischen sozialen Gruppen gekreuzt wurden: “crowds” (Menschenmassen, beispielsweise die Zuschauer eines Fußballspiels), “audience” (Zuhörerschaften, beispielsweise die Personen im Theater oder bei einer Vorlesung), “public” (Öffentlichkeit, beispielsweise Zeitungsleser oder Fernsehzuschauer), “mobs” (Mobs, wie Menschenmassen, jedoch mit wesentlich stärkerer Interaktion), “primary groups” (Primärgruppen, vertraulich), “secondary groups” (Sekundärgruppen, freiwillig) und “modern community” (die moderne Gesellschaft, eine lockere Konföderation, die aus einer engen physischen Nähe und dem Bedarf an spezialisierten Dienstleistungen entsteht).

Die folgende Tabelle zeigt die Variablen im Daten-Set, die sich aus der Klassifizierung in sieben soziale Gruppen in den Guttman-Bell-Daten ergeben, zusammen mit ihren Variablenlabeln und den Wertelabeln (Kategorien), die den Niveaus der einzelnen Variablen zugeordnet sind. Dieses Daten-Set finden Sie in *guttman.sav*. Für weitere Informationen siehe Beispieldateien in Anhang A auf S. 326. Neben der Auswahl von Variablen, die in die Berechnung der kategorialen

Hauptkomponentenanalyse eingeschlossen werden sollen, können Sie Variablen auswählen, die zur Beschriftung der Objekte in Diagrammen dienen sollen. In diesem Beispiel sind die ersten fünf Variablen in den Daten in die Analyse eingeschlossen, während "Cluster" ausschließlich als Beschriftungsvariable verwendet wird. Bei Angabe einer kategorialen Hauptkomponentenanalyse müssen Sie das optimale Skalierungsniveau für jede Analysenvariable angeben. In diesem Beispiel wird für alle Analysevariablen ein ordinales Niveau angegeben.

Tabelle 10-1
Variablen im Guttman-Bell-Daten-Set

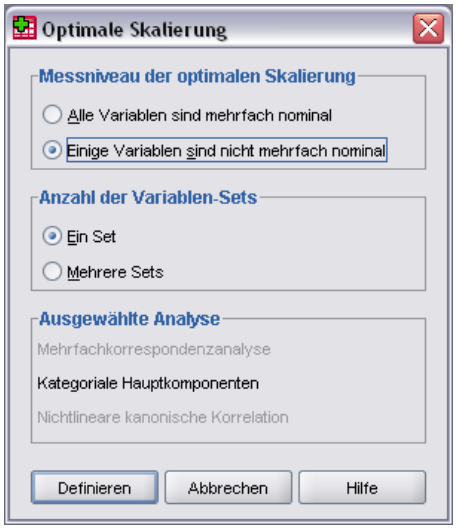
Variablenname	**Variablenlabel**	**Wertelabels**
intensit	Intensität	Sehr schwach, schwach, mittel, hoch
frequenz	Häufigkeit (Frequenz)	Schwach, nicht wiederkehrend, selten, häufig
zugehör	Zugehörigkeit	Keine, schwach, variable, hoch
entfern	Entfernung	Weit, nah
formal	Formalität	Keine Beziehung, förmlich, nicht förmlich
cluster		CR (Crowds), AU (audiences), PU (public), MB (mobs), PG (primary groups), SG (secondary groups), MC (modern community)

Durchführung der Analyse

▶ Um aus diesem Daten-Set Ausgaben für die kategoriale Hauptkomponentenanalyse zu erstellen, wählen Sie folgende Befehle aus den Menüs aus:

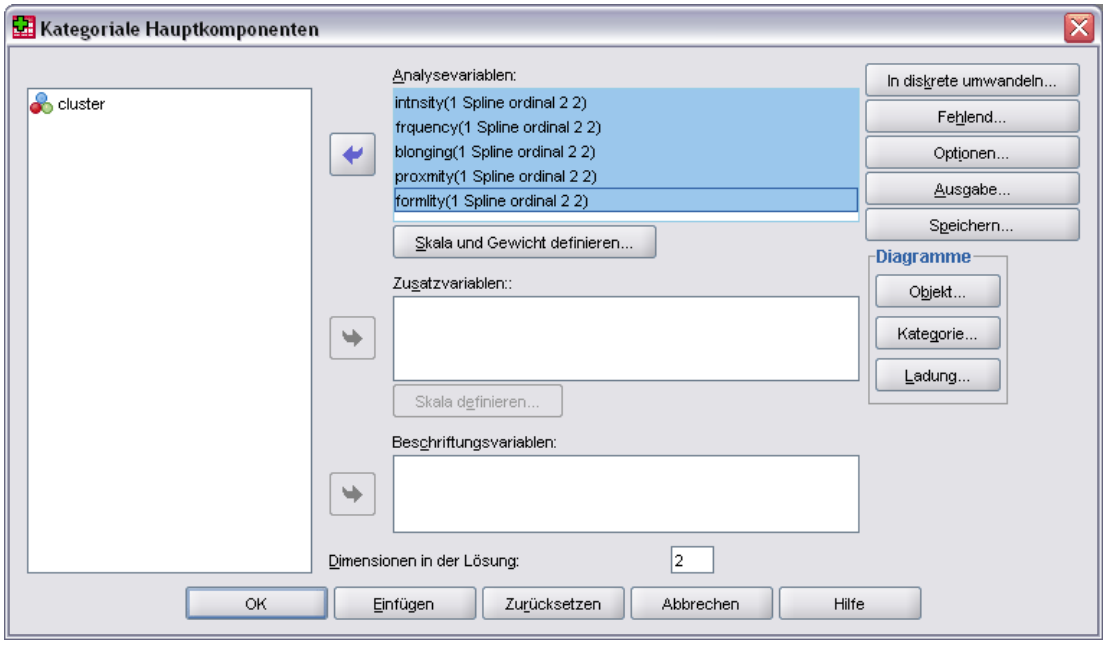
Analysieren
  Dimensionsreduktion
    Optimale Skalierung...

Abbildung 10-1
Dialogfeld "Optimale Skalierung"



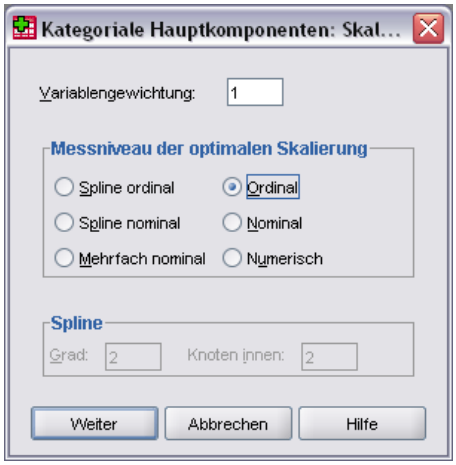
- Wählen Sie im Gruppenfeld "Messniveau der optimalen Skalierung" die Option Einige Variablen sind nicht mehrfach nominal.
- Klicken Sie auf Definieren.

Abbildung 10-2
Dialogfeld "Kategoriale Hauptkomponenten"



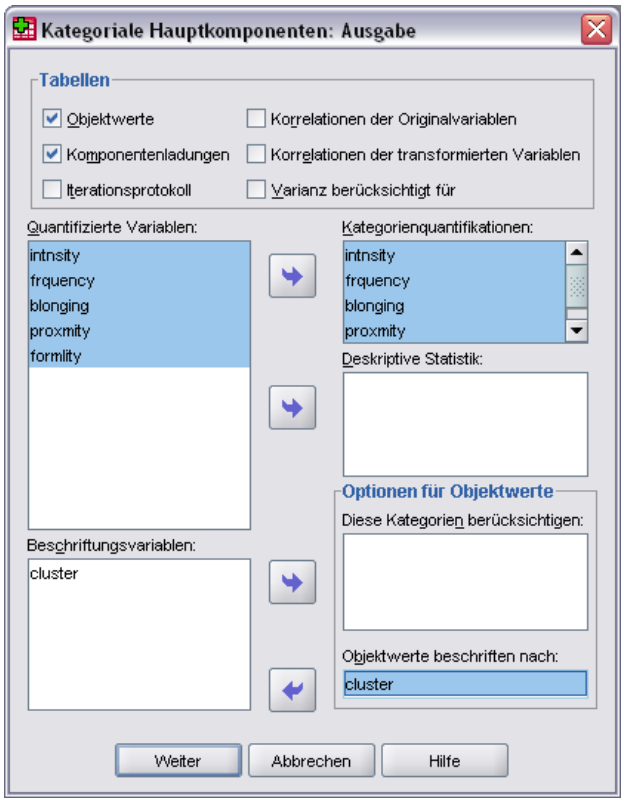
- Wählen Sie *Intensität* bis *Formalität* als Analysevariablen aus.
- Klicken Sie auf Skala und Gewicht definieren.

Abbildung 10-3
Dialogfeld "Skala und Gewichtung definieren"



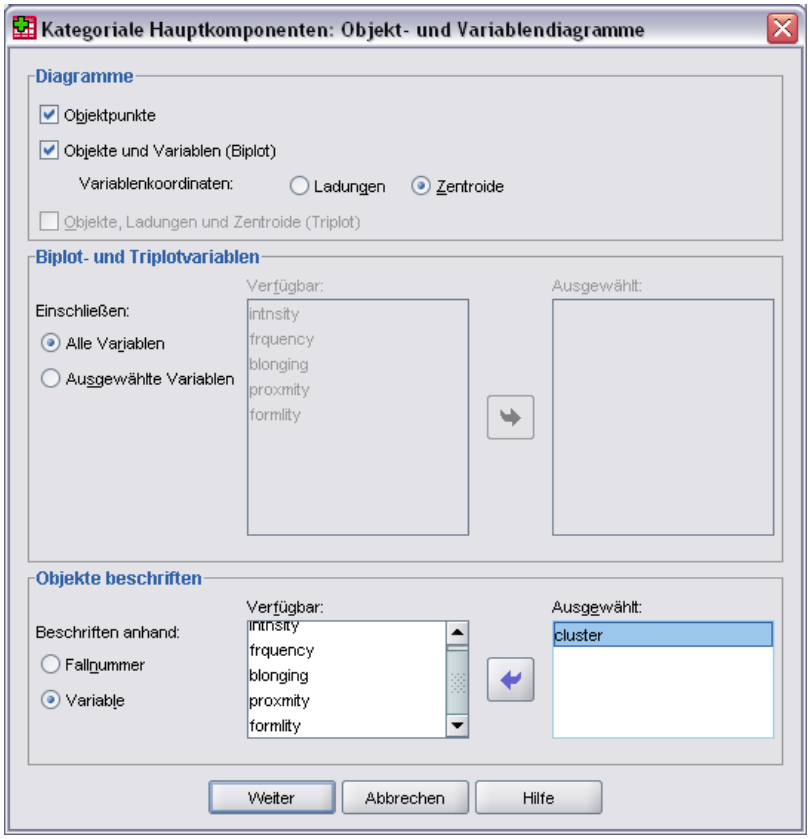
- Wählen Sie im Gruppenfeld "Messniveau der optimalen Skalierung" die Option Ordinal.
- Klicken Sie auf Weiter.
- Wählen Sie im Dialogfeld "Kategoriale Hauptkomponenten" die Variable *Cluster* als Beschriftungsvariable aus.
- Klicken Sie auf Ausgabe.

Abbildung 10-4
Dialogfeld "Ausgabe"



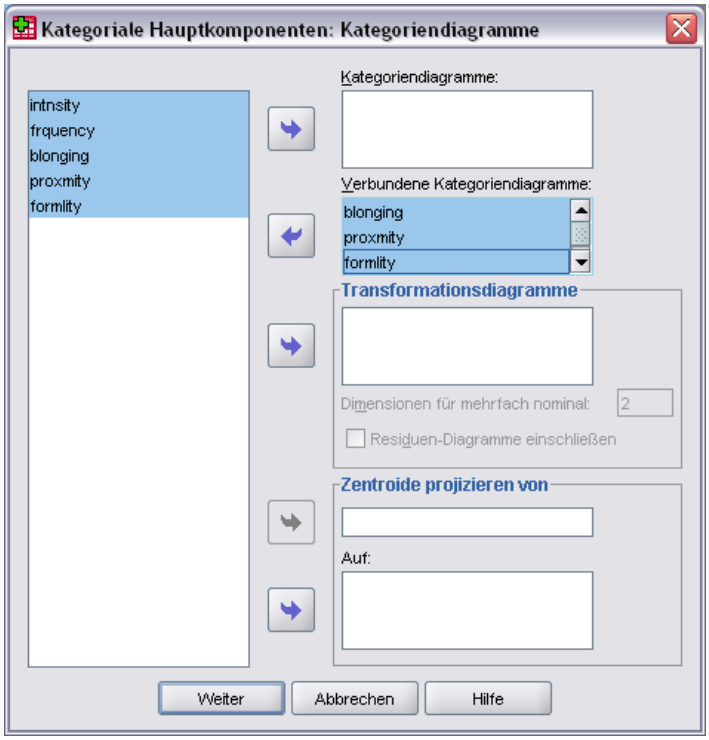
- Wählen Sie im Gruppenfeld "Tabellen" die Option Objektwerte und heben Sie die Auswahl von Korrelationen der transformierten Variablen auf.
- Aktivieren Sie die Erstellung von Kategoriequantifikationen für *intensit (Intensität)* bis *formal (Formalität)*.
- Wählen Sie "Objektwerte beschriften nach *cluster*".
- Klicken Sie auf Weiter.
- Klicken Sie im Gruppenfeld "Diagramme" des Dialogfelds "Kategoriale Hauptkomponenten" auf Objekt.

Abbildung 10-5
Dialogfeld "Objekt- und Variablendiagramme"



- Wählen Sie im Gruppenfeld "Diagramme" die Option Objekte und Variablen (Biplot).
- Wählen Sie im Gruppenfeld "Objekte beschriften" aus, dass die Objekte nach Variable beschriftet werden sollen, und wählen Sie dann *cluster* als die Variable aus, nach der die Objekte beschriftet werden sollen.
- Klicken Sie auf Weiter.
- Klicken Sie im Gruppenfeld "Diagramme" des Dialogfelds "Kategoriale Hauptkomponenten" auf Kategorie.

Abbildung 10-6
Dialogfeld "Kategoriendiagramme"



- Aktivieren Sie die Erstellung von verbundenen Kategoriendiagrammen für *intensit (Intensität)* bis *formal (Formalität)*.
- Klicken Sie auf Weiter.
- Klicken Sie im Dialogfeld "Kategoriale Hauptkomponenten" auf OK.

Anzahl der Dimensionen

Die folgenden Abbildungen zeigen einen Teil der ursprünglichen Ausgabe für die kategoriale Hauptkomponentenanalyse. Nach dem Iterationsprotokoll des Algorithmus wird die Modellzusammenfassung, einschließlich der Eigenwerte der einzelnen Dimensionen, angezeigt. Diese Eigenwerte entsprechen denen der klassischen Hauptkomponentenanalyse. Es gibt Maße dafür, wie die Varianz von den einzelnen Dimensionen berücksichtigt wird.

Abbildung 10-7
Iterationsprotokoll

	Varianz berücksichtigt		Verlustfunktion		
Iterationszahl	Gesamt	Zunahme	Gesamt	Zentroidkoordinaten	Beschränkung des Zentroids auf die Vektorkoordinaten
0[a]	4,515315	,000000	5,484685	4,075583	1,409101
31[b]	4,726009	,000008	5,273991	4,273795	1,000196

a. Iteration 0 zeigt die Statistik der Lösung mit allen als numerisch behandelten Variablen außer solchen mit dem Meßniveau der optimalen Skalierung mehrfach nominal an.

b. Der Iterationsvorgang wurde angehalten, weil der Konvergenztestwert erreicht wurde.

Abbildung 10-8
Modellzusammenfassung

		Varianz berücksichtigt	
Dimension	Cronbachs Alpha	Gesamt (Eigenwert)	% der Varianz
1	,881	3,389	67,774
2	,315	1,337	26,746
Gesamtwert	,986[a]	4,726	94,520

a. Die Summe von Cronbachs Alpha basiert auf der Summe der Eigenwerte.

Die Eigenwerte können als Anzeichen dafür verwendet werden, wie viele Dimensionen erforderlich sind. In diesem Beispiel wurde die Standardanzahl von Dimensionen, 2, verwendet. Ist das die richtige Anzahl? Als allgemeine Regel sollte der Eigenwert für eine Dimension größer als 1 sein, wenn alle Variablen entweder einzeln nominal, ordinal oder numerisch sind. Da die zweidimensionale Lösung 94,52 % der Varianz erklärt, würde eine dritte Dimension vermutlich kaum weitere Informationen ergeben.

Bei mehrfach nominalen Variablen gibt es also keine einfache Faustregel für die Bestimmung der angemessenen Anzahl an Dimensionen. Wenn die Anzahl der Variablen durch die Gesamtzahl der Kategorien minus der Anzahl der Variablen ersetzt wird, gilt die oben angegebene Regel noch immer. Doch diese Regel allein würde vermutlich mehr Dimensionen zulassen, als erforderlich sind. Bei der Auswahl der Anzahl an Dimensionen ist die nützlichste Richtlinie, die Anzahl so klein zu halten, dass sinnvolle Interpretationen möglich sind. Die Modellzusammenfassungstabelle zeigt außerdem Cronbachs Alpha (ein Reliabilitätsmaß), das durch die Prozedur maximiert wird.

Quantifizierungen

Für die einzelnen Variablen werden die Quantifizierungen, die Vektorkoordinaten und die Zentroid-Koordinaten angezeigt. Die Quantifizierungen sind die den einzelnen Kategorien zugewiesenen Werte. Die Zentroid-Koordinaten sind der Durchschnitt der Objektwerte von Objekten in derselben Kategorie. Die Vektorkoordinaten sind die Koordinaten der Kategorien, wenn sie sich auf einer Linie befinden müssen, die die Variable im Objektraum repräsentiert. Dies ist für Variablen mit ordinalem und numerischem Skalierungsniveau erforderlich.

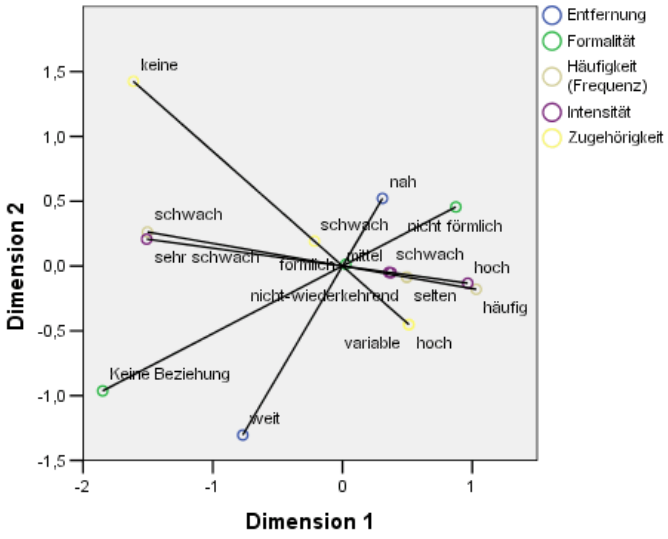
Abbildung 10-9
Quantifizierungen für die Intensität der Interaktion

Kategorie	Häufigkeit	Quantifikation	Zentroidkoordinaten Dimension 1	Zentroidkoordinaten Dimension 2	Vektorkoordinaten Dimension 1	Vektorkoordinaten Dimension 2
sehr schwach	2	-1,530	-1,496	,308	-1,510	,208
schwach	2	,362	,392	,202	,358	-,049
mittel	1	,379	,188	-1,408	,374	-,051
hoch	2	,978	1,010	,194	,965	-,133

Normalisierung mit Variablen-Prinzipal.

Ein kurzer Blick auf die Quantifizierungen im verbundenen Diagramm der Kategoriepunkte genügt, um zu sehen, dass bei manchen Variablen einige Kategorien durch die kategoriale Hauptkomponentenanalyse nicht so klar getrennt wurden, wie bei einem echt ordinalen Niveau zu erwarten gewesen wäre. Die Variablen *Intensität* und *Häufigkeit (Frequenz)* beispielsweise weisen gleiche oder fast gleiche Quantifzierungen für ihre beiden mittleren Kategorien auf. Bei dieser Art von Ergebnis sollten Sie alternative kategoriale Hauptkomponentenanalysen versuchen, vielleicht mit einigen reduzierten Kategorien oder vielleicht mit einem anderen Niveau für die Analyse, beispielsweise (mehrfach) nominal.

Abbildung 10-10
Kategoriepunkte im verbundenen Diagramm



Das verbundene Diagramm der Kategoriepunkte ähnelt dem Diagramm für die Komponentenladungen, zeigt jedoch außerdem, wo sich die Endpunkte befinden, die den niedrigsten Quantifizierungen entsprechen (z. B. *sehr schwach* für *Intensität der Interaktion* und *keine* für *Zugehörigkeit*). Die beiden Variablen, die die Interaktion messen, *Intensität* und *Häufigkeit (Frequenz)*, erscheinen sehr eng beieinander und erklären einen Großteil der Varianz in Dimension 1. *Formalität* erscheint ebenfalls nah bei *Entfernung*

Wenn Sie sich auf die Kategoriepunkte konzentrieren, können Sie die Beziehungen noch deutlicher sehen. *Intensität* und *Häufigkeit (Frequenz)* liegen nicht nur eng beieinander, sondern die Richtungen ihrer Skalen sind ähnlich; d. h. sehr schwache Intensität liegt eng bei schwacher

Häufigkeit, und häufige Interaktion liegt nahe bei hoher Intensität der Interaktion. Außerdem ist ersichtlich, dass eine enge Nähe mit einem nicht förmlichen Beziehungstyp Hand in Hand zu gehen scheint und dass eine weite Entfernung mit "Keine Beziehung" in Verbindung steht.

Objektwerte

Außerdem können Sie eine Auflistung und ein Diagramm der Objektwerte anfordern. Das Diagramm der Objektwerte kann für die Erkennung von Ausreißern, typischen Objektgruppen oder speziellen Mustern hilfreich sein.

Die Objektwertetabelle zeigt die Auflistung der Objektwerte für die Guttman-Bell-Daten, nach sozialer Gruppe beschriftet. Durch die Untersuchung der Werte für die Objektpunkte, können Sie spezielle Objekte im Diagramm identifizieren.

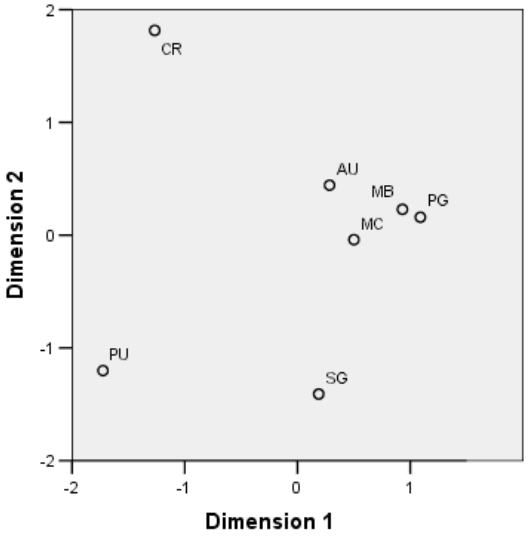
Abbildung 10-11
Objektwerte

	Dimension	
Cluster	1	2
CR	-1,266	1,816
AU	,284	,444
PU	-1,726	-1,201
MB	,931	,229
PG	1,089	,159
SG	,188	-1,408
MC	,500	-,039

Normalisierung mit Variablen-Prinzipal.

Die erste Dimension scheint *CROWDS* und *PUBLIC*, die relativ große negative Werte aufweisen, von *MOBS* und *PRIMARY GROUPS* zu trennen, die relativ große positive Werte aufweisen. In der zweiten Dimension finden sich drei Klumpen: *PUBLIC* und *SECONDARY GROUPS* mit großen negativen Werten, *CROWDS* mit großen positiven Werten und die anderen sozialen Gruppen dazwischen. Dies ist leichter zu sehen, wenn man das Diagramm der Objektwerte untersucht.

Abbildung 10-12
Objektwertediagramm

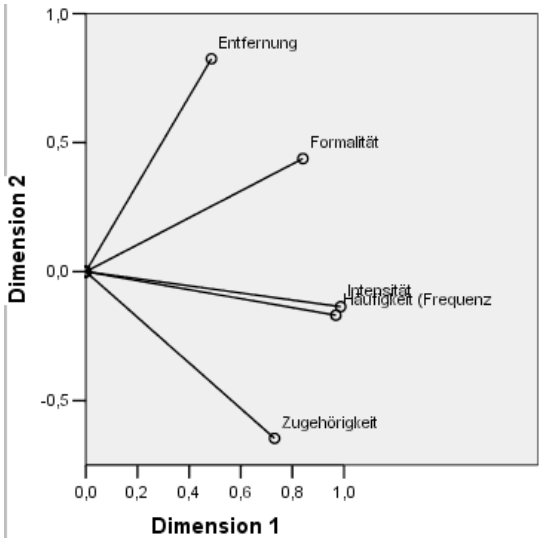


Im Diagramm sehen Sie *PUBLIC* und *SECONDARY GROUPS* unten, *CROWDS* oben und die anderen sozialen Gruppen in der Mitte. Die Untersuchung der Muster zwischen einzelnen Objekten hängt von den zusätzlichen Informationen ab, die für die Analyseeinheiten verfügbar sind. Im vorliegenden Fall kennen Sie die Klassifikation der Objekte. In anderen Fällen können Sie zur Beschriftung der Objekte Zusatzvariablen verwenden. Sie sehen außerdem, dass die kategoriale Hauptkomponentenanalyse *MOBS* nicht von *PRIMARY GROUPS* trennt. Obwohl die meisten Personen ihre Familie nicht als Mob bezeichnen würden, haben diese beiden Gruppen bei vier der fünf verwendeten Variablen dieselbe Wertung erhalten. Offensichtlich sollten die verwendeten Variablen und Kategorien auf mögliche Schwächen untersucht werden. Beispielsweise haben "hohe Interaktionsintensität" und "nicht förmliche Beziehung" für diese beiden Gruppen nicht dieselbe Bedeutung. Alternativ könnte die Verwendung einer Lösung mit mehr Dimensionen in Erwägung gezogen werden.

Komponentenladungen

Diese Abbildung zeigt das Diagramm der Komponentenladungen. Die Vektoren (Linien) sind relativ lang, was wiederum anzeigt, dass der Großteil der Varianz aller quantifizierten Variablen auf die ersten beiden Dimensionen entfällt. In der ersten Dimension haben alle Variablen hohe (positive) Komponentenladungen. Die zweite Dimension ist hauptsächlich mit den quantifizierten Variablen *Zugehörigkeit* und *Entfernung* (in entgegengesetzte Richtungen) korreliert. Das bedeutet, dass Objekte mit einem hohen negativen Wert in Dimension 2 einen hohen Wert für das Zugehörigkeitsgefühl und einen niedrigen Wert für die physische Nähe aufweisen. Die zweite Dimension macht daher einen Kontrast zwischen diesen beiden Variablen deutlich, wohingegen sie kaum in Bezug zu den quantifizierten Variablen *Interaktion* und *Häufigkeit (Frequenz)* steht.

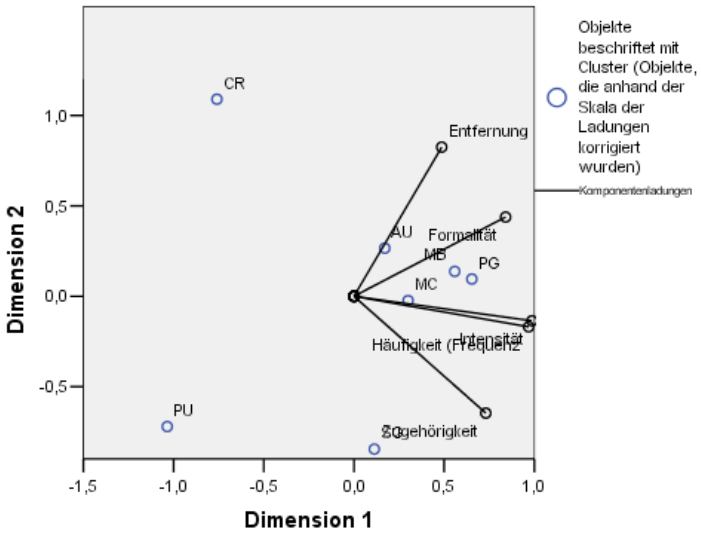
Abbildung 10-13
Komponentenladungen



Um die Relation zwischen Objekten und Variablen zu untersuchen, betrachten Sie den Biplot der Objekte und Komponentenladungen. Der Vektor einer Variablen weist in die Richtung der höchsten Kategorie der Variablen. Für *Entfernung* und *Zugehörigkeit* sind die höchsten

Kategorien *nah* bzw. *hoch*. Daher zeichnen sich *CROWDS* durch physische Nähe und kein Zugehörigkeitsgefühl aus, wohingegen *SECONDARY GROUPS* durch weite physische Entfernung und ein hohes Zugehörigkeitsgefühl charakterisiert sind.

Abbildung 10-14
Biplot



Zusätzliche Dimensionen

Durch die Erhöhung der Anzahl der Dimensionen wird der Umfang der erklärten Variation erhöht und es können eventuell Unterschiede aufgedeckt werden, die bei Lösungen mit weniger Dimensionen verborgen bleiben. Wie zuvor angemerkt, können bei zwei Dimensionen *MOBS* und *PRIMARY GROUPS* nicht voneinander getrennt werden. Bei einer Erhöhung der Dimensionalität lassen sich die beiden Gruppen jedoch möglicherweise differenzieren.

Durchführung der Analyse

- Um eine dreidimensionale Lösung zu erzielen, müssen Sie das Dialogfeld “Kategoriale Hauptkomponenten” erneut aufrufen.
- Geben Sie 3 als Anzahl der Dimensionen in der Lösung ein.
- Klicken Sie im Dialogfeld “Kategoriale Hauptkomponenten” auf OK.

Modellzusammenfassung

Abbildung 10-15
Modellzusammenfassung

Dimension	Cronbachs Alpha	Varianz berücksichtigt: Gesamt (Eigenwert)	Varianz berücksichtigt: % der Varianz
1	,885	3,424	68,480
2	-,232	,844	16,871
3	-,459	,732	14,649
Gesamtwert	1,000[a]	5,000	99,999

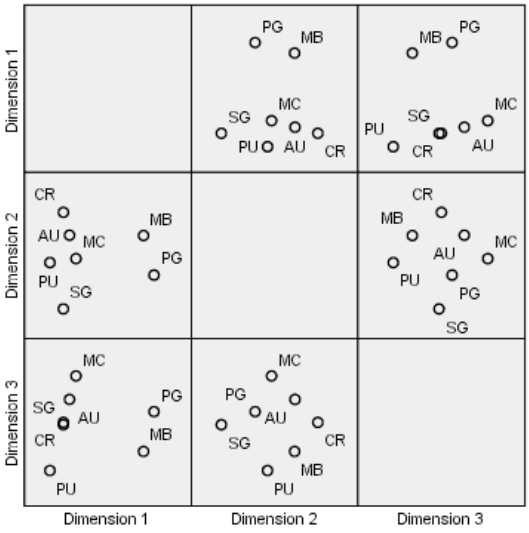
a. Die Summe von Cronbachs Alpha basiert auf der Summe der Eigenwerte.

Eine dreidimensionale Lösung weist die Eigenwerte 3,424, 0,844 und 0,732 auf, und erklärt fast die gesamte Varianz.

Objektwerte

Die Objektwerte für die dreidimensionale Lösung werden in einer Streudiagramm-Matrix grafisch dargestellt. In einer Streudiagramm-Matrix wird jede Dimension gegen jede andere Dimension in einer Reihe von zweidimensionalen Streudiagrammen geplottet. Beachten Sie: Die ersten beiden Eigenwerte in der dreidimensionalen Lösung stimmen nicht mit den Eigenwerten in der zweidimensionalen Lösung überein. Mit anderen Worten: Die Lösungen sind nicht verschachtelt. Da die Eigenwerte in den Dimensionen 2 und 3 nun kleiner als 1 sind (was zu einem negativen Wert für Cronbachs Alpha führt), sollten Sie der zweidimensionalen Lösung den Vorzug geben. Die dreidimensionale Lösung wird zu Anschauungszwecken dargestellt.

Abbildung 10-16
Streudiagramm-Matrix der Objektwerte mit drei Dimensionen



Die oberste Zeile läßt erkennen, dass die erste Dimension *PRIMARY GROUPS* und *MOBS* von den anderen Gruppen abtrennt. Beachten Sie, dass sich die Reihenfolge der Objekte entlang der vertikalen Achse in keinem der Plots in der obersten Zeile ändert. Alle diese Plots verwenden Dimension 1 als *y*-Achse.

Die mittlere Plot-Zeile ermöglicht die Interpretation von Dimension 2. Die zweite Dimension hat sich gegenüber der zweidimensionalen Lösung leicht geändert. Zuvor wies die zweite Dimension drei deutliche Klumpen auf, nun jedoch sind die Objekte weiter auf der Achse verteilt.

Die dritte Dimension unterstützt die Trennung von *MOBS* und *PRIMARY GROUPS*, die in der zweidimensionalen Lösung nicht möglich war.

Untersuchen wir die Plots "Dimension 2 versus Dimension 3" und "Dimension 1 versus Dimension 2" genauer. Auf der durch die Dimensionen 2 und 3 definierten Fläche, formen die Objekte ungefähr ein Rechteck, mit *CROWDS*, *MODERN COMMUNITY*, *SECONDARY GROUPS* und *PUBLIC* an den Eckpunkten. Auf dieser Fläche erscheinen *MOBS* und *PRIMARY GROUPS* als konvexe Kombinationen von *PUBLIC-CROWDS* bzw. *SECONDARY GROUPS-MODERN COMMUNITY*. Sie werden jedoch, wie bereits erwähnt, entlang Dimension 1 von den anderen Gruppen getrennt. *AUDIENCES* ist entlang Dimension 1 nicht von den anderen Gruppen getrennt und scheint eine Kombination von *CROWDS* und *MODERN COMMUNITY* zu sein.

Komponentenladungen

Abbildung 10-17
Dreidimensionale Komponentenladungen

	Dimension		
	1	2	3
Intensität	,980	-,005	-,201
Häufigkeit (Frequenz)	,521	-,643	,561
Zugehörigkeit	,980	-,002	-,197
Entfernung	,519	,656	,549
Formalität	,981	,004	-,193

Normalisierung mit Variablen-Prinzipal.

Zu wissen, wie die Objekte getrennt sind, bedeutet noch nicht, dass klar ist, welche Variablen welchen Dimensionen entsprechen. Dies wird mithilfe der Komponentenladungen ermittelt. Die erste Dimension entspricht in erster Linie *Zugehörigkeit*, *Intensität* und *Formalität*; die zweite Dimension trennt *Häufigkeit (Frequenz)* und *Entfernung*; und die dritte Dimension trennt diese Elemente von den anderen.

Beispiel: Symptomatologie von Eßstörungen

Eßstörungen sind schwächende Krankheiten, die mit Störungen im Eßverhalten, schwerer Abweichung des Körperbilds und einer Besessenheit von Gewicht einhergehen, die Geist und Körper gleichzeitig betrifft. Jedes Jahr sind Millionen von Menschen betroffen, wobei Heranwachsende besonders gefährdet sind. Eine Behandlung ist möglich und am hilfreichsten, wenn der Zustand frühzeitig erkannt wird.

Mediziner können versuchen, Eßstörungen durch eine psychologische und medizinische Bewertung zu diagnostizieren. Es kann sich jedoch als schwierig erweisen, einen Patienten einer von mehreren Klassen von Eßstörungen zuzuordnen, da es keine standardisierte Sympomatologie anorektischen/bulimischen Verhaltens gibt. Gibt es Symptome, anhand deren die Patienten deutlich in die vier Gruppen unterteilt werden können? Welche Symptome haben sie gemeinsam?

Um diese Fragen zu beantworten, führten Wissenschaftler (Van der Ham, Meulman, Van Strien, als auch Van Engeland, 1997) eine Studie an 55 Jugendlichen mit bekannten Eßstörungen durch, wie in der folgenden Tabelle gezeigt.

Tabelle 10-2
Patientendiagnose

Diagnose	**Anzahl der Patienten**
Anorexia nervosa	25
Anorexia mit Bulimia nervosa	9
Bulimia nervosa nach Anorexia	14
Atypische Eßstörung	7
Gesamt	55

Jeder Patient wurde vier Mal über einen Zeitraum von vier Jahren untersucht, es fanden also insgesamt 220 Beobachtungen statt. Bei jeder Beobachtung wurden die Patienten anhand der 16 Symptome bewertet, die in der folgenden Tabelle dargestellt sind. Die Symptomwerte fehlen für Patient 71 zum Zeitpunkt 2, Patient 76 zum Zeitpunkt 2 und Patient 47 zum Zeitpunkt 3, wodurch 217 gültige Beobachtungen verbleiben. Diese Daten finden Sie in *anorectic.sav*.Für weitere Informationen siehe Beispieldateien in Anhang A auf S. 326.

Tabelle 10-3
Modifizierte Morgan-Russell-Unterskalen zur Messung des Wohlbefindens

Variablen-name	**Variablenlabel**	**Unteres Ende (Wert 1)**	**Oberes Ende (Wert 3 oder 4)**
weight	Body weight (Körpergewicht)	Outside normal range (Außerhalb des normalen Bereichs)	Normal (Normal)
mens	Menstruation (Menstruation)	Amenorrhea (Amenorrhö)	Regular periods (Regelmäßige Periode)
fast	Restriction of food intake (fasting) (Einschränkung der Nahrungsaufname (Fasten))	Less than 1200 calories (Weniger als 1200 Kalorien)	Normal/regular meals (Normale/regelmäßige Mahlzeiten)
binge	Binge eating (Freßattacken)	Greater than once a week (Mehr als einmal pro Woche)	No bingeing (Keine Freßattacken)
vomit	Vomiting (Erbrechen)	Greater than once a week (Mehr als einmal pro Woche)	No vomiting (Kein Erbrechen)
purge	Purging (Abführmittel)	Greater than once a week (Mehr als einmal pro Woche)	No purging (Keine Abführmittel)
hyper	Hyperactivity (Hyperaktivität)	Not able to be at rest (Kann nicht zur Ruhe kommen)	No hyperactivity (Keine Hyperaktivität)

Variablen-name	Variablenlabel	Unteres Ende (Wert 1)	Oberes Ende (Wert 3 oder 4)
fami	Family relations (Familienbeziehungen)	Poor (Schlecht)	Good (Gut)
eman	Emancipation from family (Emanzipation von der Familie)	Very dependent (Sehr abhängig)	Adequate (Angemessen)
frie	Friends (Freunde)	No good friends (Keine guten Freunde)	Two or more good friends (Mindestens zwei gute Freunde)
school	School/employment record (Schulische/berufliche Leistungen)	Stopped school/work (Schule/Arbeit abgebrochen)	Moderate to good record (Mittel bis gut)
satt	Sexual attitude (Sexuelle Einstellung)	Inadequate (Unangemessen)	Adequate (Angemessen)
sbeh	Sexual behavior (Sexualverhalten)	Inadequate (Unangemessen)	Can enjoy sex (Kann Sex genießen)
mood	Mental state (mood) (Psychische Verfassung (Stimmung))	Very depressed (Sehr niedergeschlagen)	Normal (Normal)
preo	Preoccupation with food and weight (Beschäftigung mit Essen und Gewicht)	Complete (Vollständig)	No preoccupation (Keine Beschäftigung)
body	Body perception (Körperwahrnehmung)	Disturbed (Gestört)	Normal (Normal)

Für diese Situation ist die Hauptkomponentenanalyse ideal, da der Zweck der Studie darin besteht, die Beziehungen zwischen den Symptomen und den verschiedenen Klassen von Eßstörungen zu bestimmen. Außerdem ist die kategoriale Hauptkomponentenanalyse wahrscheinlich nützlicher als die klassische Hauptkomponentenanalyse, da die Wertung für die Symptome auf einer ordinalen Skala erfolgt.

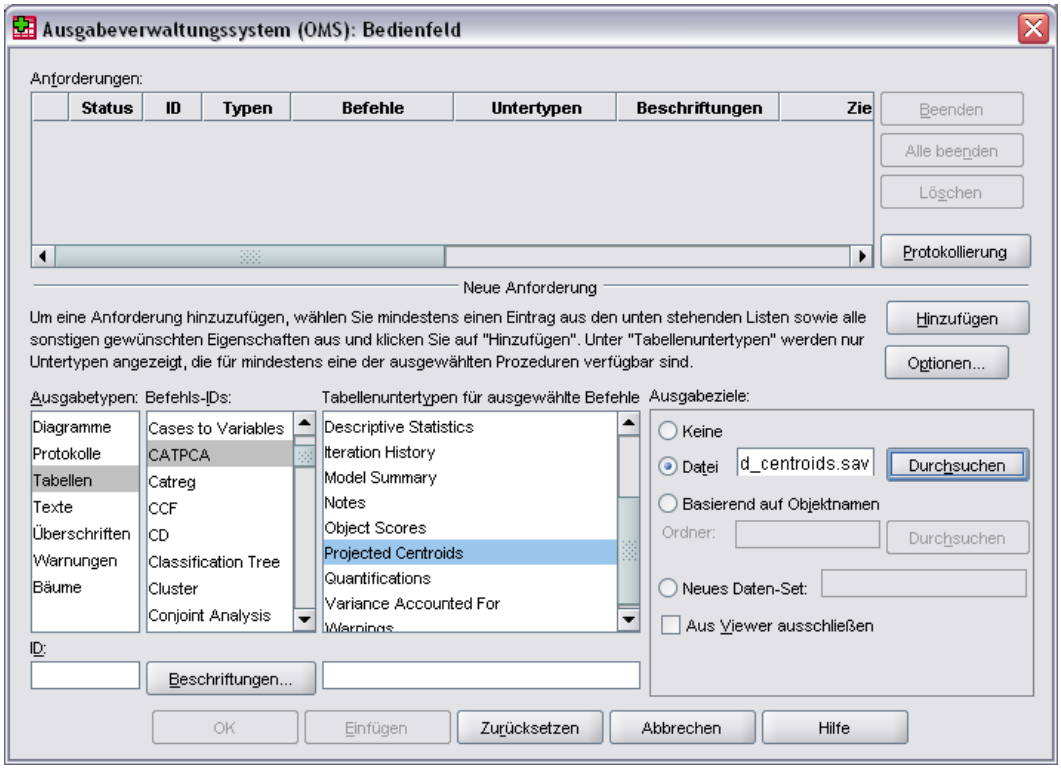
Durchführung der Analyse

Um die Struktur des Krankheitsverlaufs für die einzelnen Diagnosen ordnungsgemäß besser untersuchen zu können, sollten Sie die Ergebnisse der Tabelle der projizierten Zentroide als Daten für Streudiagramme verfügbar machen. Sie können dies mithilfe des Ausgabeverwaltungssystems (OMS) erreichen.

► Um eine OMS-Anforderung zu starten, wählen Sie die folgenden Befehle aus den Menüs aus:

Extras
  OMS-Steuerung...

Abbildung 10-18
Systemsteuerung des Ausgabeverwaltungssystems

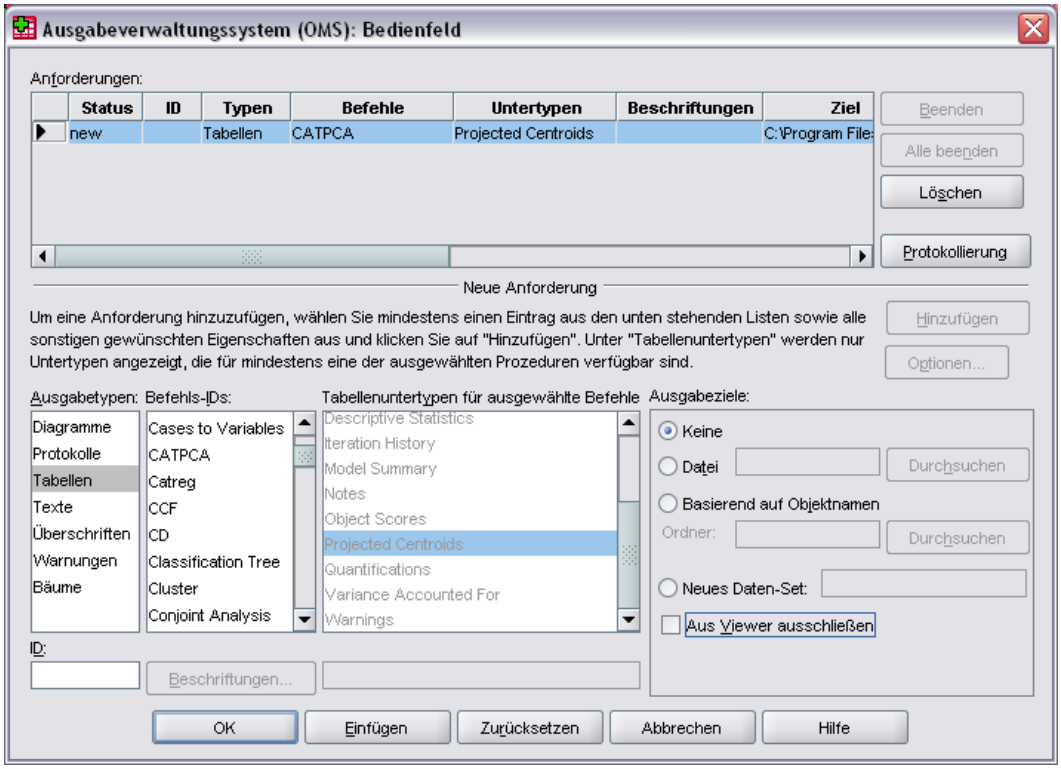


- Wählen Sie Tabellen als Ausgabetyp.
- Wählen Sie CATPCA als Befehl.
- Wählen Sie Projected Centroids als Tabellentyp.
- Wählen Sie im Gruppenfeld “Ausgabeziele” die Option Datei und geben Sie als Dateinamen projected_centroids.sav ein.
- Klicken Sie auf Optionen.

Abbildung 10-19
Dialogfeld "Optionen"

- Wählen Sie als Ausgabeformat SPSS-Datendatei.
- Geben Sie TableNumber_1 als Tabellennummernvariable ein.
- Klicken Sie auf Weiter.

Abbildung 10-20
Systemsteuerung des Ausgabeverwaltungssystems

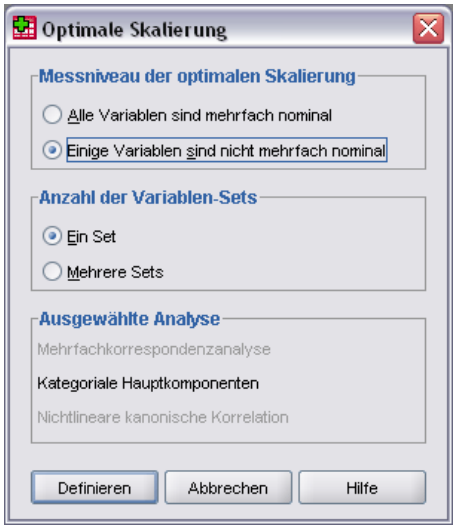


- Klicken Sie auf Hinzufügen.
- Klicken Sie auf OK und dann noch einmal auf OK, um die OMS-Sitzung zu bestätigen.

  Das Ausgabeverwaltungssystem (OMS) ist nun so eingerichtet, dass es die Ergebnisse der Tabelle der projizierten Zentroide in die Datei *projected_centroids.sav* schreibt.
- Um aus diesem Daten-Set Ausgaben für die kategoriale Hauptkomponentenanalyse zu erstellen, wählen Sie folgende Befehle aus den Menüs aus:

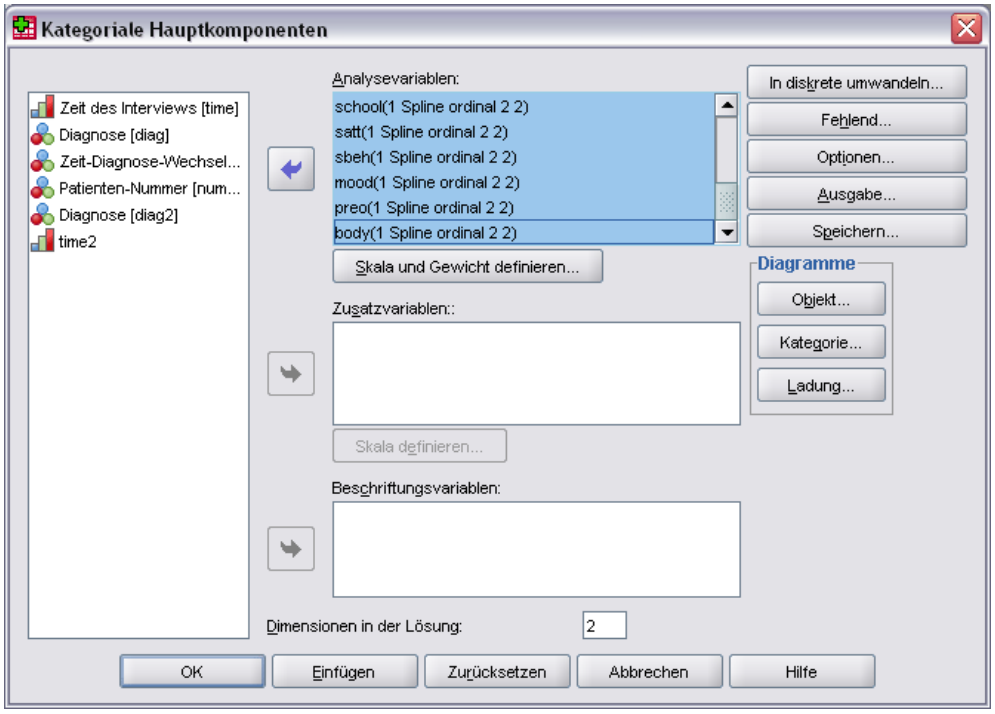
  Analysieren
  Dimensionsreduktion
  Optimale Skalierung...

Abbildung 10-21
Dialogfeld "Optimale Skalierung"



- Wählen Sie im Gruppenfeld "Messniveau der optimalen Skalierung" die Option Einige Variablen sind nicht mehrfach nominal.
- Klicken Sie auf Definieren.

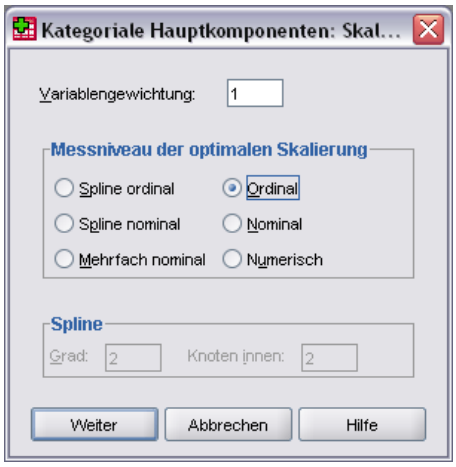
Abbildung 10-22
Dialogfeld "Kategoriale Hauptkomponenten"



- Wählen Sie *Body Weight* (Körpergewicht) bis *Body perception* (Körperwahrnehmung) als Analysevariablen aus.

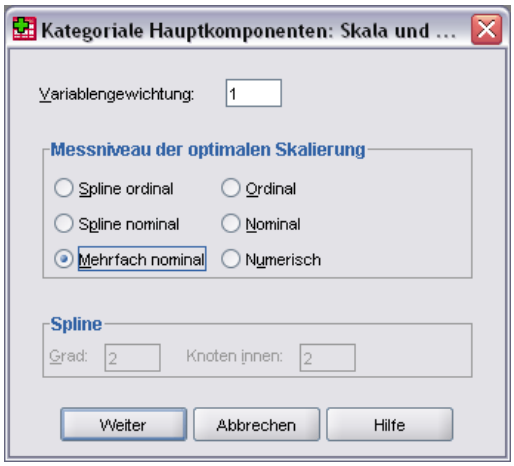
- Klicken Sie auf Skala und Gewicht definieren.

Abbildung 10-23
Dialogfeld "Skala und Gewichtung definieren"



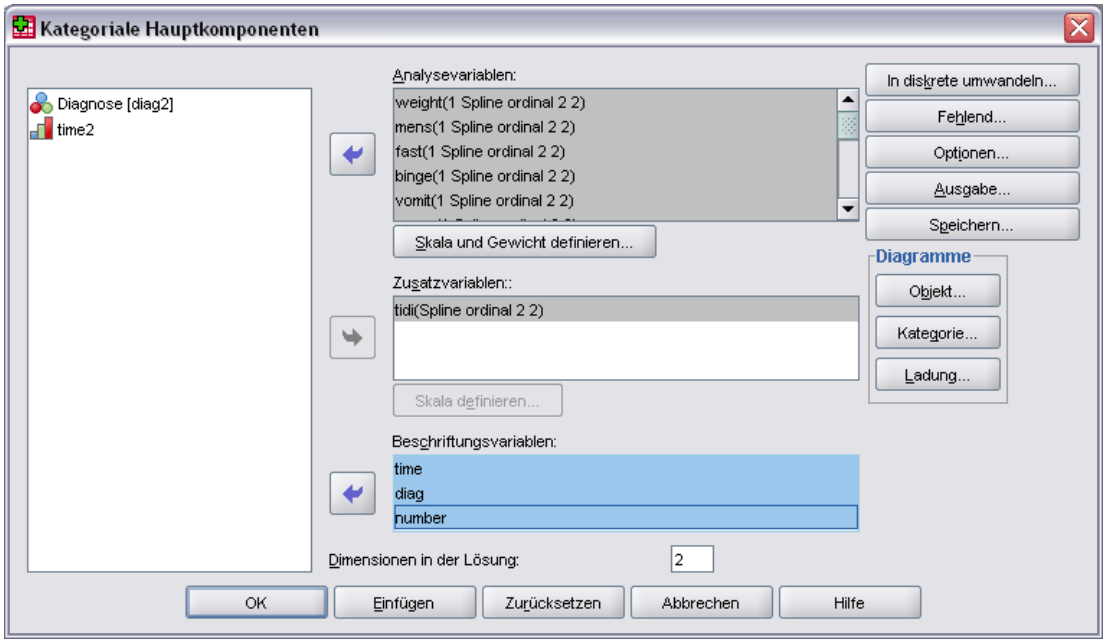
- Wählen Sie als Messniveau der optimalen Skalierung die Option Ordinal aus.
- Klicken Sie auf Weiter.
- Wählen Sie *Time/diagnosis interaction* (Zeit/Diagnose-Interaktion) als Zusatzvariable aus und klicken Sie im Dialogfeld "Kategoriale Hauptkomponenten" auf Skala definieren.

Abbildung 10-24
Dialogfeld "Skala definieren"



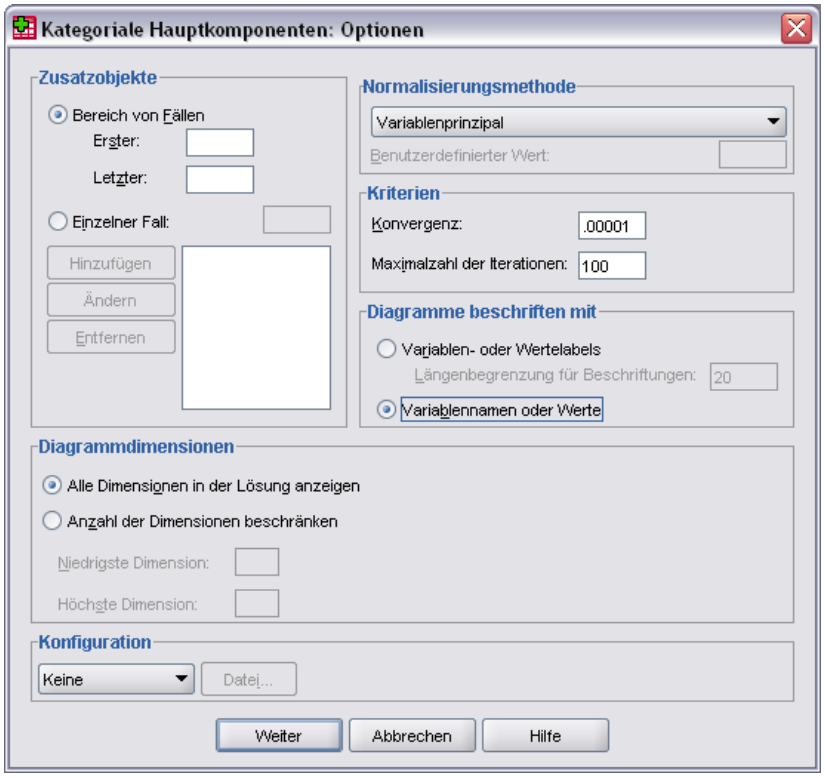
- Wählen Sie Mehrfach nominal als Messniveau der optimalen Skalierung.
- Klicken Sie auf Weiter.

Abbildung 10-25
Dialogfeld "Kategoriale Hauptkomponenten"



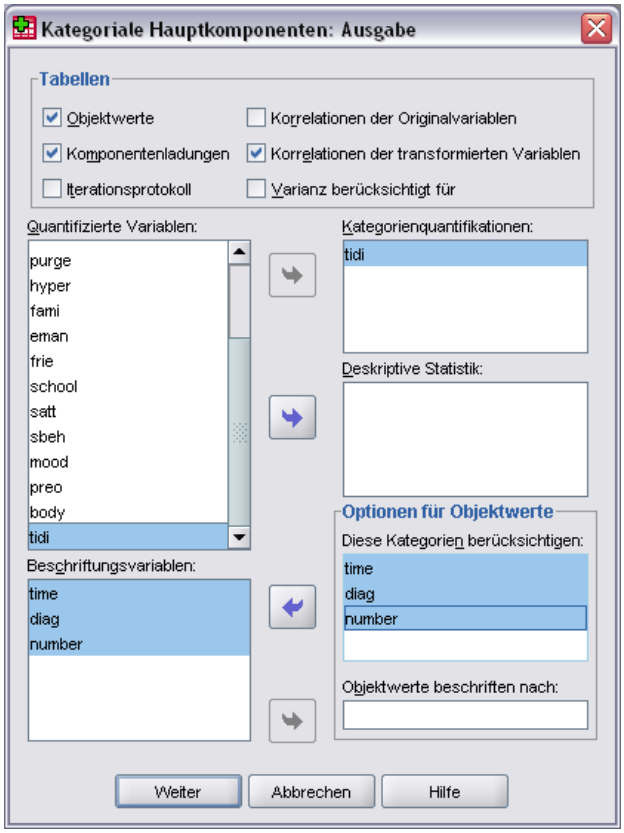
- Wählen Sie *Time of interview* (Zeitpunkt der Befragung) bis *Patient Number* (Patientennummer) als Beschriftungsvariablen aus.
- Klicken Sie auf Optionen.

Abbildung 10-26
Dialogfeld "Optionen"



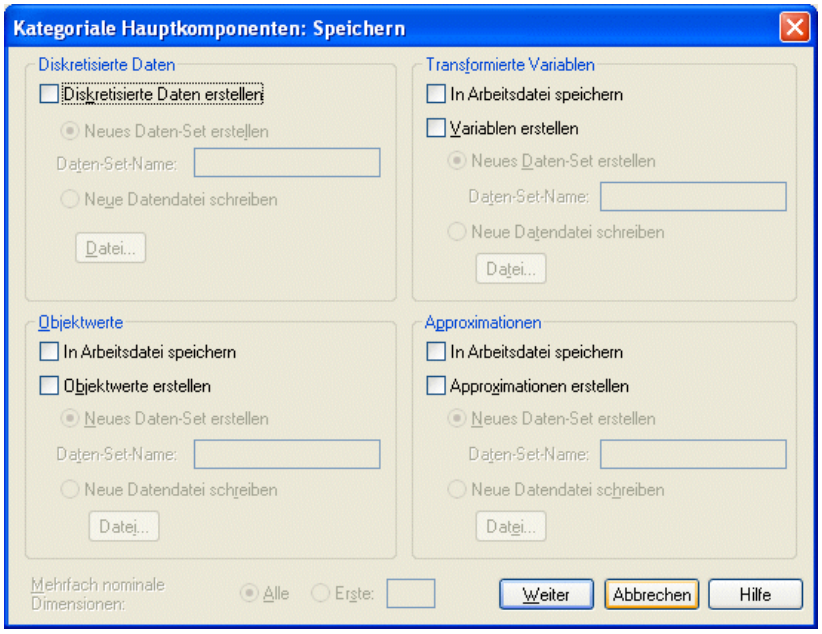
- Wählen Sie Variablennamen oder Werte als Option für die Diagrammbeschriftung.
- Klicken Sie auf Weiter.
- Klicken Sie im Dialogfeld "Kategoriale Hauptkomponenten" auf Ausgabe.

Abbildung 10-27
Dialogfeld "Ausgabe"



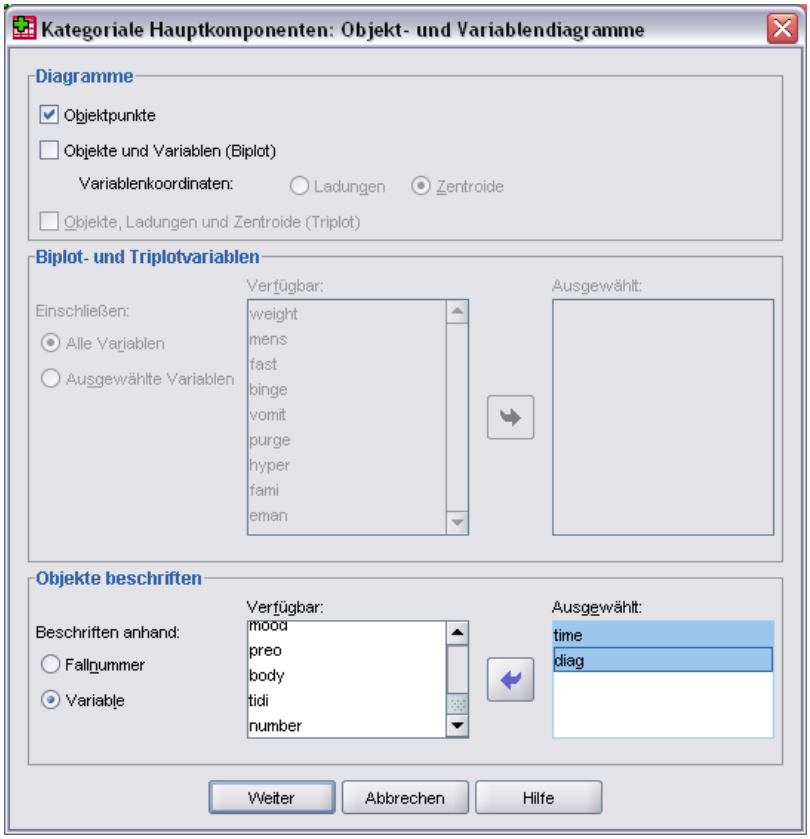
- Wählen Sie im Gruppenfeld "Tabelle" die Option Objektwerte.
- Fordern Sie die Kategoriequantifikationen für *tidi* an.
- Wählen Sie aus, dass die Kategorien *time*, *diag* und *number* berücksichtigt werden sollen.
- Klicken Sie auf Weiter.
- Klicken Sie im Dialogfeld "Kategoriale Hauptkomponenten" auf Speichern.

Abbildung 10-28
Dialogfeld "Speichern"



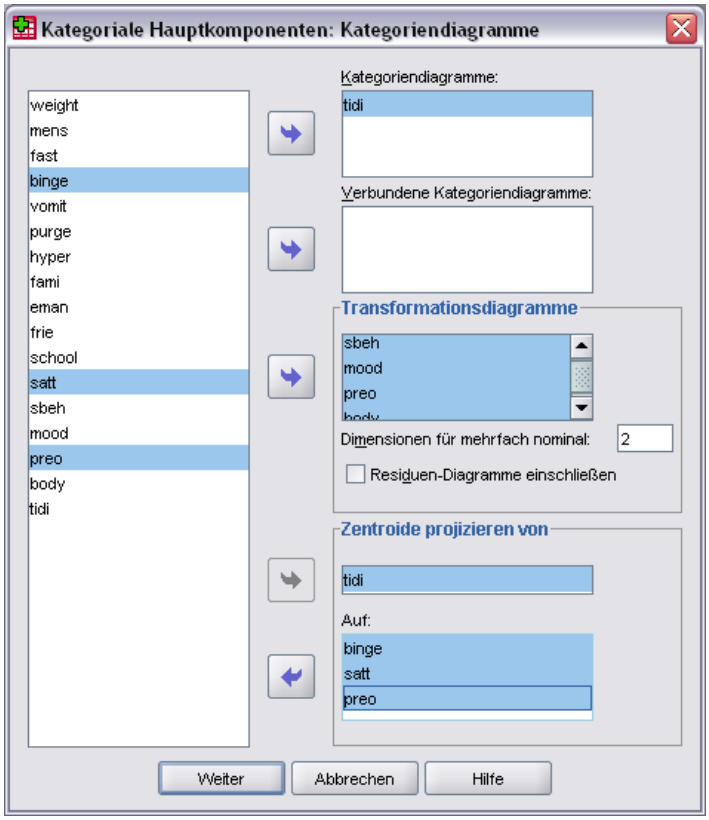
- Wählen Sie im Gruppenfeld "Transformierte Variablen" die Option In Arbeitsdatei speichern aus.
- Klicken Sie auf Weiter.
- Klicken Sie im Dialogfeld "Kategoriale Hauptkomponenten" auf Objekt.

Abbildung 10-29
Dialogfeld "Objekt- und Variablendiagramme"



- Wählen Sie aus, dass Objekte nach Variable beschriftet werden sollen.
- Wählen Sie *time* und *diag* als die Variablen aus, nach denen die Objekte beschriftet werden sollen.
- Klicken Sie auf Weiter.
- Klicken Sie im Dialogfeld "Kategoriale Hauptkomponenten" auf Kategorie.

Abbildung 10-30
Dialogfeld "Kategoriendiagramme"

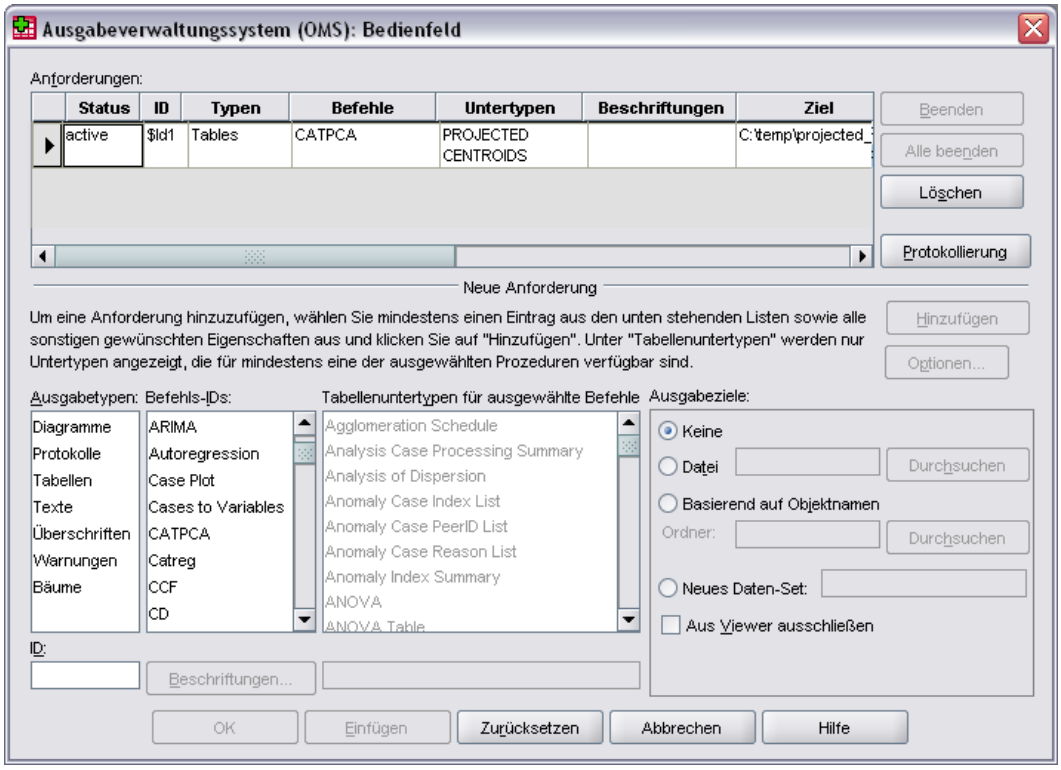


- Fordern Sie Kategoriendiagramme für *tidi* an.
- Fordern Sie Transformationsdiagramme für *weight* bis *body* an.
- Wählen Sie aus, dass die Zentroide von *tidi* auf *binge*, *satt* und *preo* projiziert werden sollen.
- Klicken Sie auf Weiter.
- Klicken Sie im Dialogfeld "Kategoriale Hauptkomponenten" auf OK.

Die Prozedur ergibt Werte für die Subjekte (mit Mittelwert 0 und Einheitsvarianz) und Quantifizierungen der Kategorien, die die mittlere quadrierte Korrelation der Subjektwerte und der transformierten Variablen maximieren. Bei der vorliegenden Analyse wurden die Kategoriequantifikationen eingeschränkt, sodass sie die ordinalen Informationen widerspiegeln.

Schließlich müssen Sie die OMS-Anforderung beenden, um die Informationen aus der Tabelle der projizierten Zentroide in die Datei *projected_centroids.sav* zu schreiben. Rufen Sie die OMS-Systemsteuerung erneut auf.

Abbildung 10-31
Systemsteuerung des Ausgabeverwaltungssystems

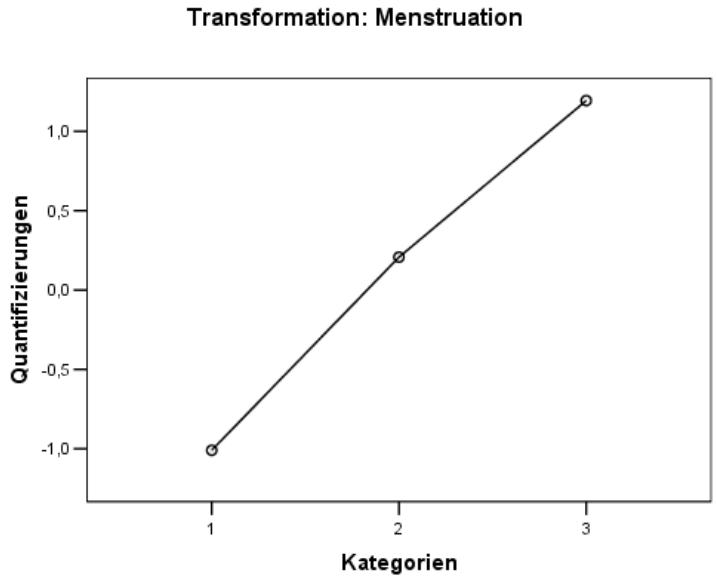


- Klicken Sie auf Beenden.
- Klicken Sie auf OK und dann zur Bestätigung noch einmal auf OK.

Transformationsdiagramme

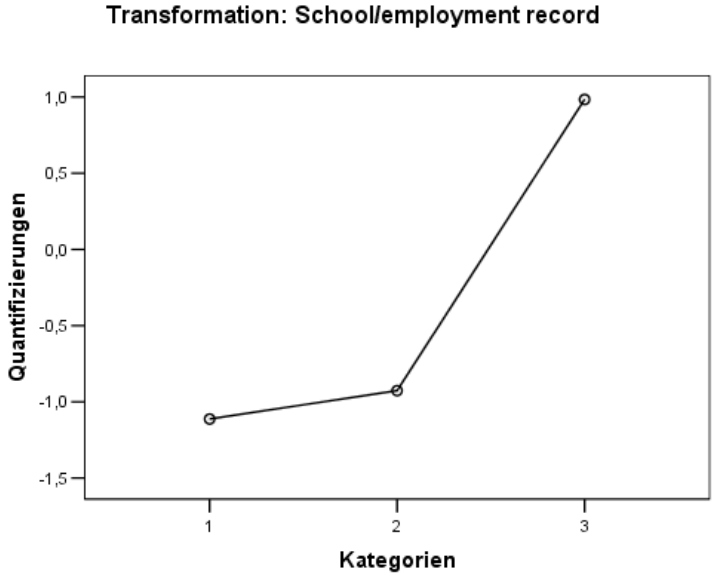
Die Transformationsdiagramme zeigen die ursprüngliche Kategorienummer auf den horizontalen Achsen an; die vertikalen Achsen zeigen die optimalen Quantifizierungen.

Abbildung 10-32
Transformationsdiagramm für die Menstruation



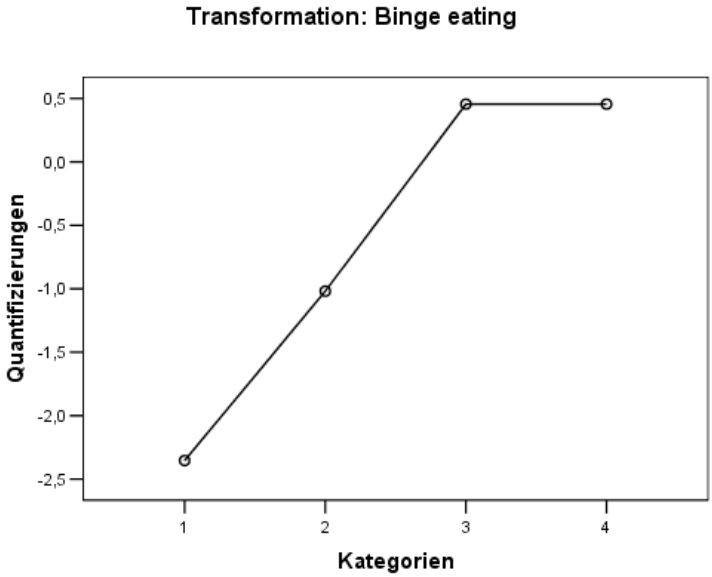
Einige Variablen, wie *Menstruation* (Menstruation), ergaben fast lineare Transformationen. Daher können Sie sie in dieser Analyse als numerisch interpretieren.

Abbildung 10-33
Transformationsdiagramm für "School/employment record" (schulische/berufliche Leistungen)



Die Quantifizierungen für andere Variablen, wie *School/employment record* (schulische/berufliche Leistungen), ergaben keine linearen Transformationen und sollten auf dem ordinalen Skalierungsniveau interpretiert werden. Der Unterschied zwischen der zweiten und der dritten Kategorie ist wesentlich wichtiger als der Unterschied zwischen der ersten und der zweiten Kategorie.

Abbildung 10-34
Transformationsdiagramm für "Binge eating" (Freßattacken)



Ein interessanter Fall ergibt sich bei den Quantifizierungen für *Binge eating* (Freßattacken). Die erzielte Transformation ist linear für die Kategorien 1 bis 3, doch die quantifizierten Werte für die Kategorien 3 und 4 sind gleich. Dieses Ergebnis zeigt, dass die Werte für 3 und 4 zu keiner Unterscheidung zwischen den Patienten führen, und deutet darauf hin, dass Sie das numerische Skalierungsniveau in einer Zwei-Komponenten-Lösung verwenden könnten, indem Sie die Werte für 4 als Werte für 3 aufzeichnen.

Modellzusammenfassung

Abbildung 10-35
Modellzusammenfassung

Modellzusammenfassung

Dimension	Cronbachs Alpha	Varianz berücksichtigt: Gesamt (Eigenwert)	Varianz berücksichtigt: % der Varianz
1	,874	5,550	34,690
2	,522	1,957	12,234
Gesamtwert	,925[a]	7,508	46,924

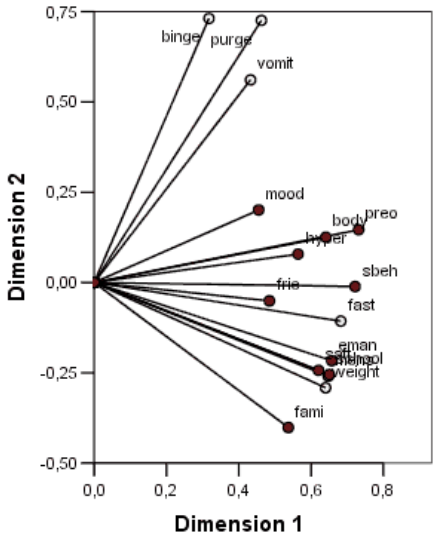
a. Die Summe von Cronbachs Alpha basiert auf der Summe der Eigenwerte.

Um zu sehen, wie gut das Modell an die Daten angepaßt ist, untersuchen wir die Modellzusammenfassung. Etwa 47 % der gesamten Varianz wird durch das Zwei-Komponenten-Modell erklärt, 35 % durch die erste Dimension und 12 % durch die zweite. Also wird fast die Hälfte der Variabilität auf der Ebene der einzelnen Objekte durch das Zwei-Komponenten-Modell erklärt.

Komponentenladungen

Zu Beginn der Interpretation der beiden Dimensionen der Lösung, betrachten wir die Komponentenladungen. Alle Variablen weisen eine positive Komponentenladung in der ersten Dimension auf, was bedeutet, dass ein gemeinsamer Faktor vorliegt, der mit allen Variablen positiv korreliert.

Abbildung 10-36
Komponentenladungsdiagramme



Die zweite Dimension trennt die Variablen. Die Variablen *Binge eating* (Freßattacken), *Vomiting* (Erbrechen) und *Purging* (Abführmittel) bilden ein Bündel mit großen positiven Ladungen in der zweiten Dimension. Diese Symptome werden üblicherweise als charakteristisch für bulimisches Verhalten betrachtet.

Die Variablen *Emancipation from family* (Emanzipation von der Familie), *School/employment record* (schulische/berufliche Leistungen), *Sexual attitude* (sexuelle Einstellung), *Body weight* (Körpergewicht) und *Menstruation* (Menstruation) bilden ein weiteres Bündel, in das auch *Restriction of food intake (fasting)* (Einschränkung der Nahrungsaufnahme (Fasten)) und *Family relations* (Familienbeziehungen) aufgenommen werden kann, da die Vektoren dieser Variablen eng beim Hauptklumpen liegen und diese Variablen als Symptome für Anorexie gelten (Fasten, Gewicht, Menstruation) oder psychosozialer Natur sind (Emanzipation, schulische/berufliche Leistungen, sexuelle Einstellung, Familienbeziehungen). Die Vektoren dieses Bündels sind rechtwinklig (lotrecht) zu den Vektoren von "binge", "vomit" und "purge", was bedeutet, dass diese Gruppe von Variablen nicht mit der Gruppe der Bulimie-Variablen korreliert ist.

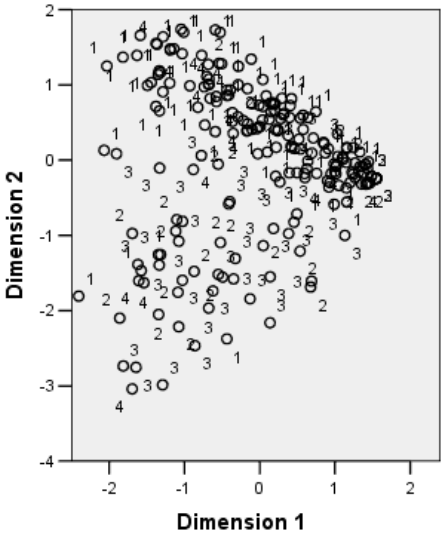
Die Variablen *Friends* (Freunde), *Mental state (mood)* (psychische Verfassung (Stimmung)) und *Hyperactivity* (Hyperaktivität) scheinen nicht sehr gut in die Lösung zu passen. Sie können dies am Diagramm sehen, indem Sie die Länge der einzelnen Vektoren beobachten. Die Vektorlänge einer Variablen entspricht ihrer Anpassung und die genannten Variablen haben die kürzesten Vektoren. Auf der Grundlage einer Zwei-Komponenten-Lösung würden Sie diese Variablen vermutlich aus einer potentiellen Symptomatologie für Eßstörungen herausnehmen. In einer höherdimensionierten Lösung können sie jedoch besser passen.

Die Variablen *Sexual behavior* (Sexualverhalten), *Preoccupation with food and weight* (Beschäftigung mit Essen und Gewicht) und *Body perception* (Körperwahrnehmung) bilden eine andere theoretische Symptomgruppe, die damit zu tun hat, wie der Patient seinen Körper erlebt. Diese Variablen sind zwar mit den beiden im rechten Winkel zueinander stehenden Variablenbündeln korreliert, weisen jedoch ziemlich lange Vektoren auf und sind eng mit der ersten Dimension verbunden, weshalb sie einige nützliche Informationen über den "gemeinsamen" Faktor liefern können.

Objektwerte

Die folgende Abbildung zeigt ein Diagramm der Objektwerte, bei denen dei Subjekte mit ihrer jeweiligen Diagnosekategorie beschriftet sind.

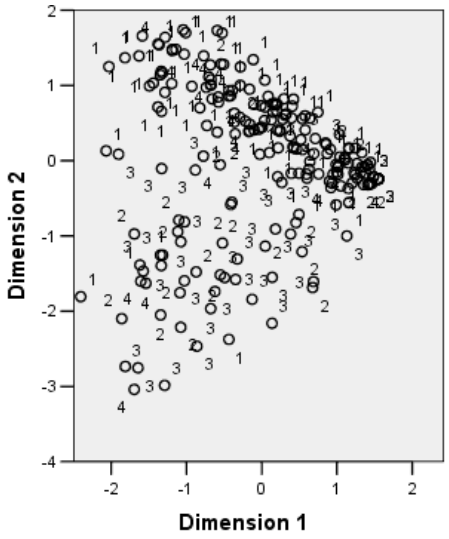
Abbildung 10-37
Objektwertediagramm, mit Diagnose beschriftet



Dieses Diagramm erleichtert nicht die Interpretation der ersten Dimension, weil darin die Patienten nicht nach Diagnose getrennt werden. Es gibt jedoch Informationen über die zweite Dimension. Patienten mit Anorexie (1) und Patienten, die an einer atypischen Eßstörung leiden (4), bilden eine Gruppe, die sich oberhalb der Personen befindet, die eine Art von Bulimie aufweisen (2 und 3). Die zweite Dimension trennt also bulimische Patienten von den anderen, wie auch im vorherigen Abschnitt deutlich wurde (die Variablen im Bulimie-Bündel weisen große positive Komponentenladungen in der zweiten Dimension auf). Dies ergibt einen Sinn, wenn man bedenkt, dass die Komponentenladungen der Symptome, die traditionell mit Bulimie assoziiert werden, große Werte in der zweiten Dimension aufweisen.

Die folgende Abbildung zeigt ein Diagramm der Objektwerte, bei denen die Subjekte mit dem jeweiligen Diagnosezeitpunkt beschriftet sind.

Abbildung 10-38
Objektwerte mit "Time of interview" (Zeitpunkt der Befragung) beschriftet



Wenn die Objektwerte nach Zeitpunkt beschriftet werden, zeigt sich, dass die erste Dimension eine Relation zur Zeit aufweist, da die Diagnosezeitpunkte für 1 größtenteils links und die anderen eher weiter rechts liegen. Sie können die Zeitpunkte in diesem Diagramm verbinden, indem Sie die Objektwerte speichern und ein Streudiagramm mit den Werten (Scores) von Dimension 1 auf der *x*- und den Scores von Dimension 2 auf der *y*-Achse erstellen und anschließend die Markierungen unter Verwendung der Patientennummern festlegen.

Ein Vergleich des nach Zeitpunkt beschrifteten Objektwertediagramms mit dem nach Diagnose beschrifteten Diagramm kann Aufschluss über ungewöhnliche Objekte gewähren. So gibt es in dem nach Zeitpunkt beschrifteten Diagramm einen Patienten, dessen Diagnose zum Zeitpunkt 4 links von allen anderen Punkten im Diagramm liegt. Dies ist ungewöhnlich, da der allgemeine Trend der Punkte dahin geht, dass die späteren Zeitpunkte weiter rechts liegen. Interessanterweise weist dieser Punkt, der zeitlich deplaziert zu sein scheint, auch eine ungewöhnliche Diagnose auf: der Patient ist ein Anorektiker, dessen Werte im Bulimie-Klumpen liegen. Wenn wir die Tabelle der Objektwerte betrachten, ergibt sich, dass es sich hierbei um Patient 43 handelt, bei dem Anorexia nervosa diagnostiziert wurde. Die Objektwerte diese Patienten finden Sie in folgender Tabelle:

Tabelle 10-4
Objektwerte für Patient 43

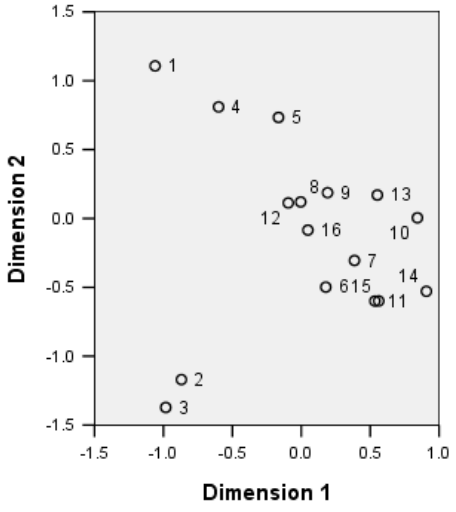
Zeit	Dimension 1	Dimension 2
1	–2.031	1.250
2	–2.067	0.131
3	–1.575	–1.467
4	–2.405	–1.807

Die Werte des Patienten zum Zeitpunkt 1 sind prototypisch für Anorektiker, mit einem hohen negativen Wert in Dimension 1, der einem schlechten Körperbild entspricht, und dem positiven Wert in Dimension 2, der Anorexie-Symptomen bzw. schlechtem psychosozialem Verhalten entspricht. Jedoch ist anders als bei der Mehrzahl der Patienten kaum oder gar kein Fortschritt in Dimension 1 zu verzeichnen. In Dimension 2 gibt es offensichtlich einen gewissen Fortschritt in Richtung "normal" (um den Wert 0, zwischen anorektischem und bulimischem Verhalten), aber dann tritt eine Verschiebung hin zu bulimischen Symptomen ein.

Untersuchung der Strukturen im Krankheitsverlauf

Um weitere Informationen dazu herauszufinden, wie die beiden Dimensionen mit den vier Diagnosekategorien und den vier Zeitpunkten in Bezug standen, wurde eine Zusatzvariable, *Time/diagnosis interaction* (Zeit/Diagnose-Interaktion) erstellt, indem eine Kreuzklassifikation der vier Kategorien von *Patient diagnosis* (Patientendiagnose) und der vier Kategorien von *Time of interview* (Zeitpunkt der Befragung) durchgeführt wurde. Daher weist *Time/diagnosis interaction* (Zeit/Diagnose-Interaktion) 16 Kategorien auf, wobei die erste Kategorie für die Anorexia-nervosa-Patienten bei ihrem ersten Besuch steht. Die fünfte Kategorie steht für die Anorexia-nervosa-Patienten bei Zeitpunkt 2 usw. bis zur 16. Kategorie, die für die Patienten mit atypischer Essstörung bei Zeitpunkt 4 steht. Durch die Verwendung der Zusatzvariablen *Time/diagnosis interaction* (Zeit/Diagnose-Interaktion) wird die Untersuchung des Krankheitsverlaufs für die verschiedenen Gruppen im Laufe der Zeit möglich. Die Variable erhielt ein mehrfach nominales Skalierungsniveau. Die Kategoriepunkte werden in der folgenden Abbildung angezeigt.

Abbildung 10-39
Kategoriepunkte für "Time/diagnosis interaction" (Zeit/Diagnose-Interaktion)

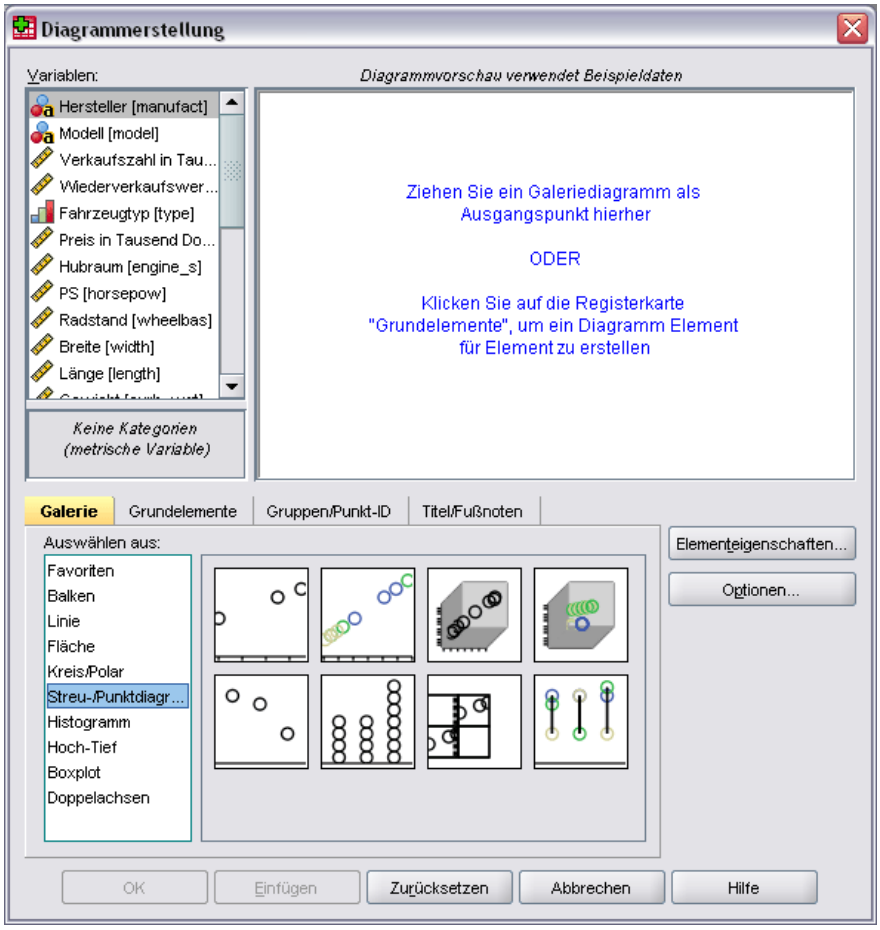


Ein Teil der Struktur ist aus diesem Diagramm ersichtlich: Die Diagnosekategorien bei Zeitpunkt 1 trennen in der zweiten Dimension klar Anorexia nervosa und atypische Eßstörung von Anorexia nervosa mit Bulimia nervosa und Bulimia nervosa nach Anorexia nervosa. Danach ist es etwas schwieriger, die Muster zu sehen.

Sie können die Muster jedoch einfacher sichtbar machen, indem Sie ein Streudiagramm auf der Grundlage der Quantifizierungen erstellen. Wählen Sie hierzu die folgenden Befehle aus den Menüs aus:

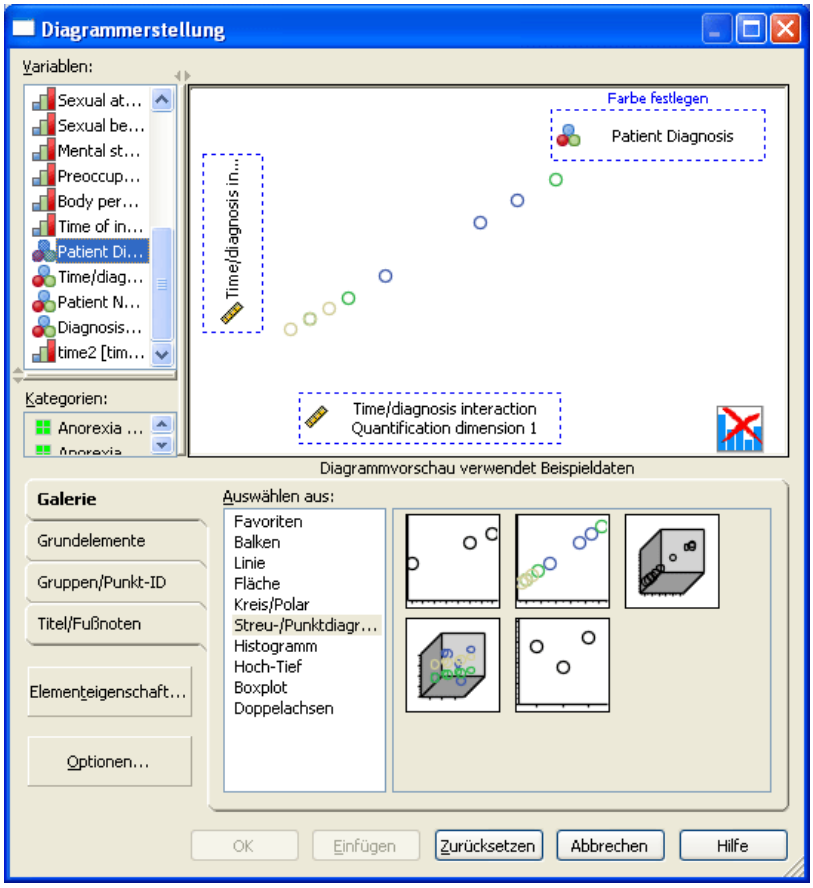
Grafiken
Diagrammerstellung...

Abbildung 10-40
Galerie "Streu-/Punktdiagramm"



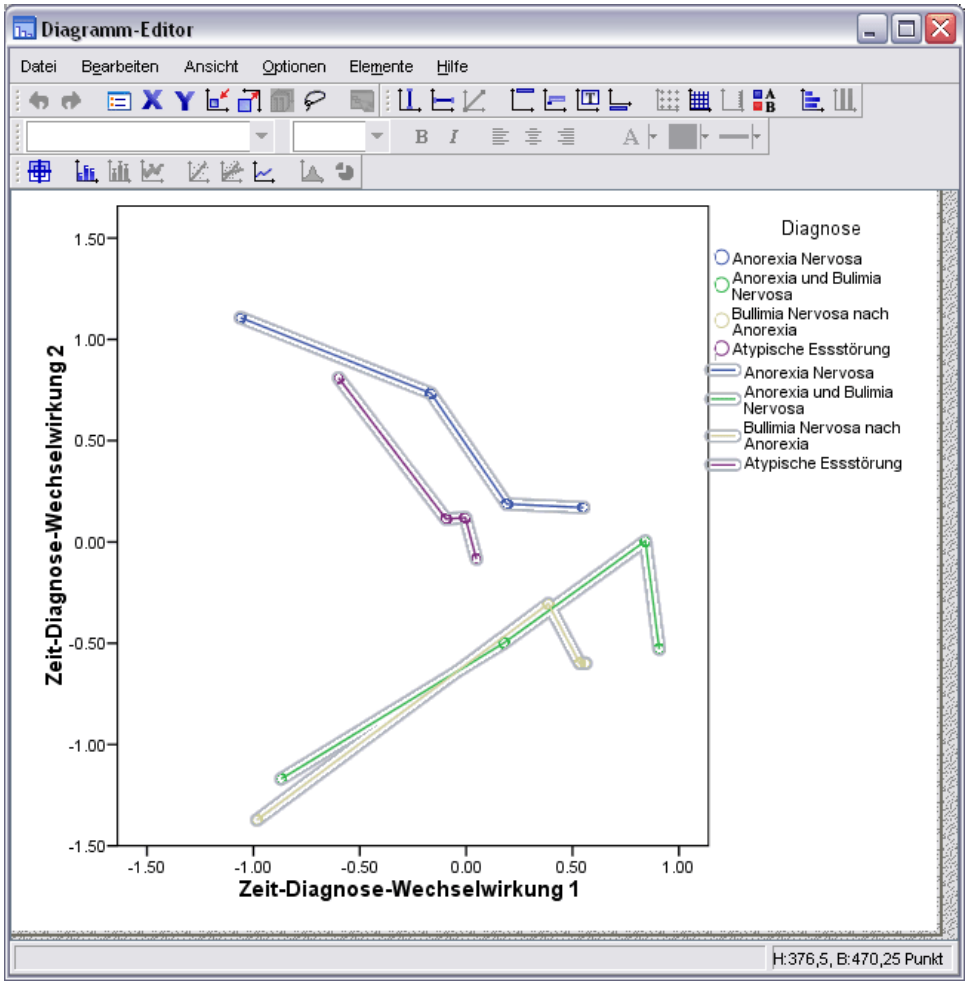
▶ Wählen Sie die Galerie Streu-/Punktdiagramm aus und wählen Sie die Option "Gruppiertes Streudiagramm".

Abbildung 10-41
Diagrammerstellung



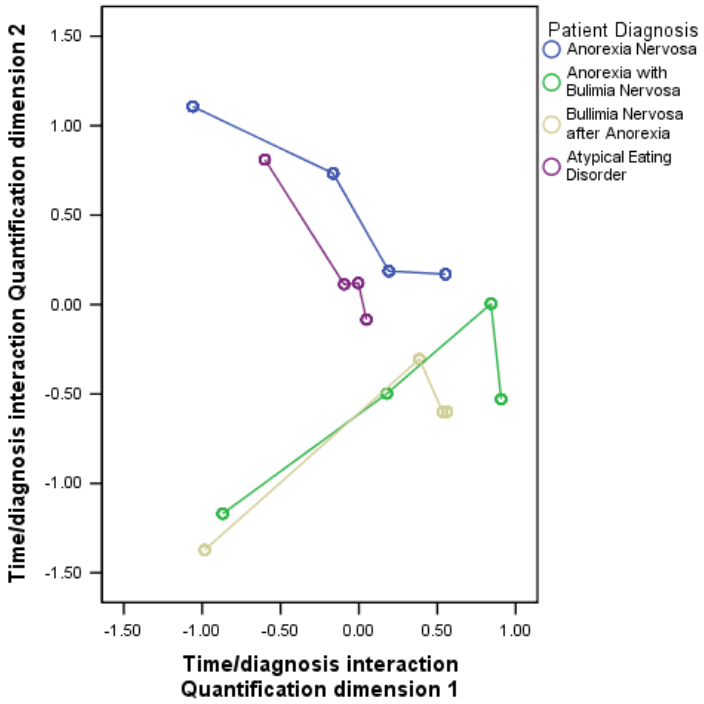
- Wählen Sie *Time/diagnosis interaction Quantification dimension 2* (Zeit/Diagnose-Interaktion, Quantifizierung Dimension 2) als *y*-Achsen-Variable und *Time/diagnosis interaction Quantification dimension 1* (Zeit/Diagnose-Interaktion, Quantifizierung Dimension 1) als *x*-Achsen-Variable.
- Wählen Sie aus, dass die Farbe anhand von *Patient Diagnosis* (Patientendiagnose) festgelegt werden soll.
- Klicken Sie auf OK.

Abbildung 10-42
Strukturen des Krankheitsverlaufs



- Doppelklicken Sie dann zur Verbindung der Punkte auf das Diagramm und klicken Sie im Diagramm-Editor auf das Tool zum Hinzufügen von Interpolationslinien.
- Schließen Sie den Diagramm-Editor.

Abbildung 10-43
Strukturen des Krankheitsverlaufs



Durch Verbindung der Kategoriepunkte für die einzelnen Diagnosekategorien im Verlauf der Zeit ergeben sich Muster, die sofort nahelegen, dass die erste Dimension eine Relation zur Zeit und die zweite Dimension eine Relation zur Diagnose aufweist, wie wir zuvor aus den Objektwertediagrammen ermittelt haben.

Dieses Diagramm zeigt jedoch zusätzlich, dass sich die Krankheiten im Verlauf der Zeit tendenziell aneinander angleichen. Außerdem ist der Fortschritt bei allen Gruppen zwischen den Zeitpunkten 1 und 2 am größten; die anorektischen Patienten zeigen noch einen gewissen weiteren Fortschritt von 2 bis 3, bei den anderen Gruppen ist jedoch kaum weiterer Fortschritt zu verzeichnen.

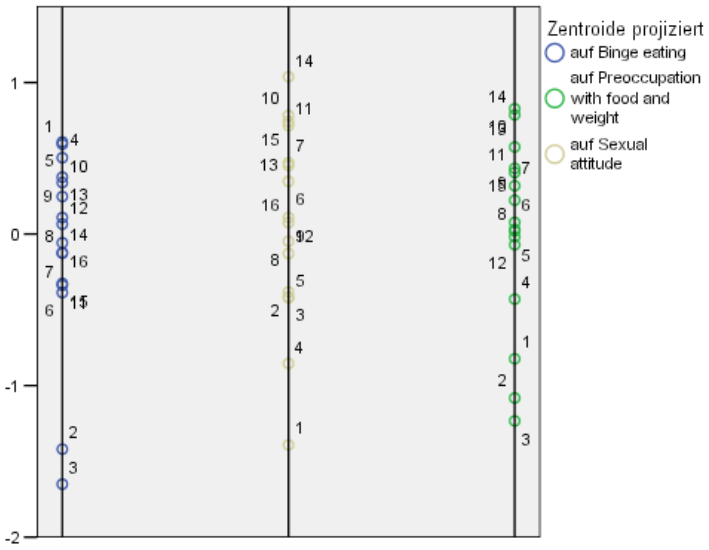
Unterschiedliche Entwicklung für ausgewählte Variablen

Aus jedem der durch die Komponentenladungen identifizierten Symptombündel wurde eine Variable als "Repräsentant" des Bündels ausgewählt. "Binge eating" (Freßattacken) wurde aus dem Bulimie-Bündel ausgewählt, "Sexual attitude" (sexuelle Einstellung) aus dem Bündel "Anorexie/psychosoziales Verhalten" und "Preoccupation with food and weight" (Beschäftigung mit Essen und Gewicht) aus dem dritten Bündel.

Um die möglichen unterschiedlichen Krankheitsverläufe zu untersuchen, wurden die Projektionen von *Time/diagnosis interaction* (Zeit/Diagnose-Interaktion) auf *Binge eating* (Freßattacken), *Sexual attitude* (sexuelle Einstellung) und *Preoccupation with food and weight*

(Beschäftigung mit Essen und Gewicht) berechnet und in der folgenden Abbildung grafisch dargestellt.

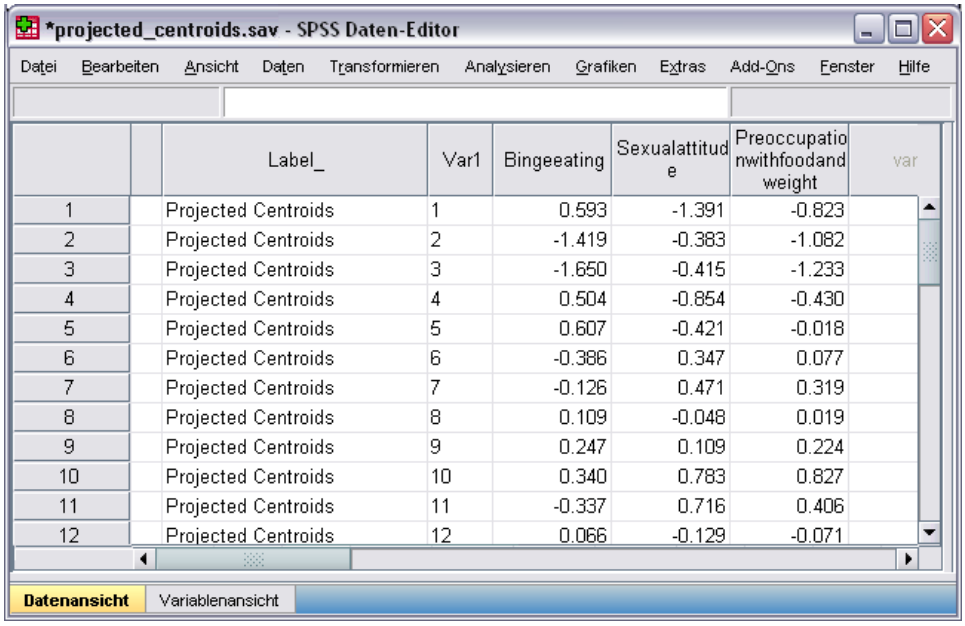
Abbildung 10-44
Projizierte Zentroide von "Time/diagnosis interaction" (Zeit/Diagnose-Interaktion) auf "Binge eating" (Freßattacken), "Sexual attitude" (sexuelle Einstellung) und "Preoccupation with food and weight" (Beschäftigung mit Essen und Gewicht)



Dieses Diagramm zeigt, dass beim ersten Zeitpunkt das Symptom "Binge eating" (Freßattacken) die Bulimie-Patienten (2 und 3) von den anderen (1 und 4) trennt; die sexuelle Einstellung trennt die anorektischen und atypischen Patienten (1 und 2) von den anderen (2 und 3); "Preoccupation with food and weight" (Beschäftigung mit Essen und Gewicht) bietet keine Trennung der Patienten. Bei vielen Anwendungen würde dieses Diagramm ausreichen, um die Beziehung zwischen den Symptomen und der Diagnose zu beschreiben, aufgrund der höheren Komplexität durch die Zeitpunkte, wird das Bild jedoch unklar.

Um diese Projektionen im Verlauf der Zeit zu betrachten, müssen Sie in der Lage sein, den Inhalt der Tabelle der projizierten Zentroide grafisch darzustellen. Dies wird durch die OMS-Anforderung ermöglicht, mit der diese Informationen in der Datei *projected_centroids.sav* gespeichert wurden.

Abbildung 10-45
Projected_centroids.sav


*projected_centroids.sav - SPSS Daten-Editor

Datei Bearbeiten Ansicht Daten Transformieren Analysieren Grafiken Extras Add-Ons Fenster Hilfe

	Label_	Var1	Bingeeating	Sexualattitude	Preoccupationwithfoodandweight	var
1	Projected Centroids	1	0.593	-1.391	-0.823	
2	Projected Centroids	2	-1.419	-0.383	-1.082	
3	Projected Centroids	3	-1.650	-0.415	-1.233	
4	Projected Centroids	4	0.504	-0.854	-0.430	
5	Projected Centroids	5	0.607	-0.421	-0.018	
6	Projected Centroids	6	-0.386	0.347	0.077	
7	Projected Centroids	7	-0.126	0.471	0.319	
8	Projected Centroids	8	0.109	-0.048	0.019	
9	Projected Centroids	9	0.247	0.109	0.224	
10	Projected Centroids	10	0.340	0.783	0.827	
11	Projected Centroids	11	-0.337	0.716	0.406	
12	Projected Centroids	12	0.066	-0.129	-0.071	

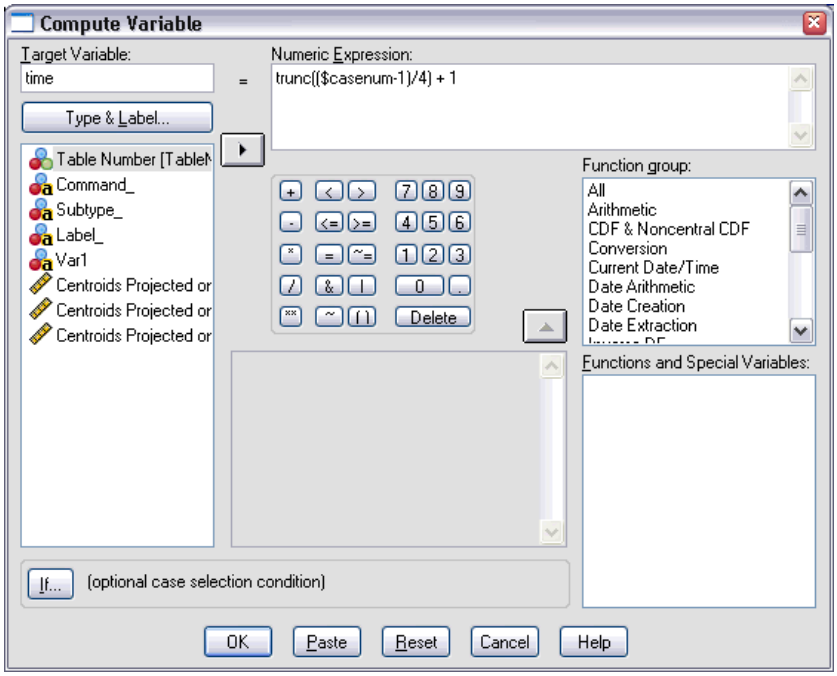
Datenansicht Variablenansicht

Die Variablen *Bingeeating*, *Sexualattitude* und *Preoccupationwithfoodandweight* enthalten die Werte der Zentroide, die auf die einzelnen relevanten Symptome projiziert wurden. Die Fallnummer (1 bis 16) entspricht “Time/diagnosis interaction” (Zeit/Diagnose-Interaktion). Sie müssen neue Variablen berechnen, die die Zeit- und Diagnosewerte trennen.

▶ Wählen Sie die folgenden Befehle aus den Menüs aus:

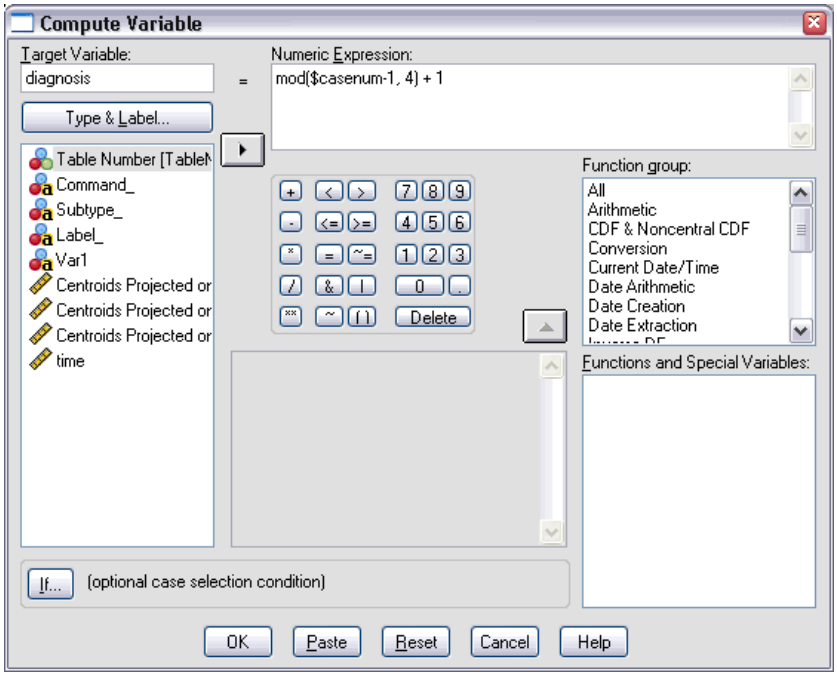
Transformieren
  Variable berechnen...

Abbildung 10-46
Dialogfeld "Variable berechnen"



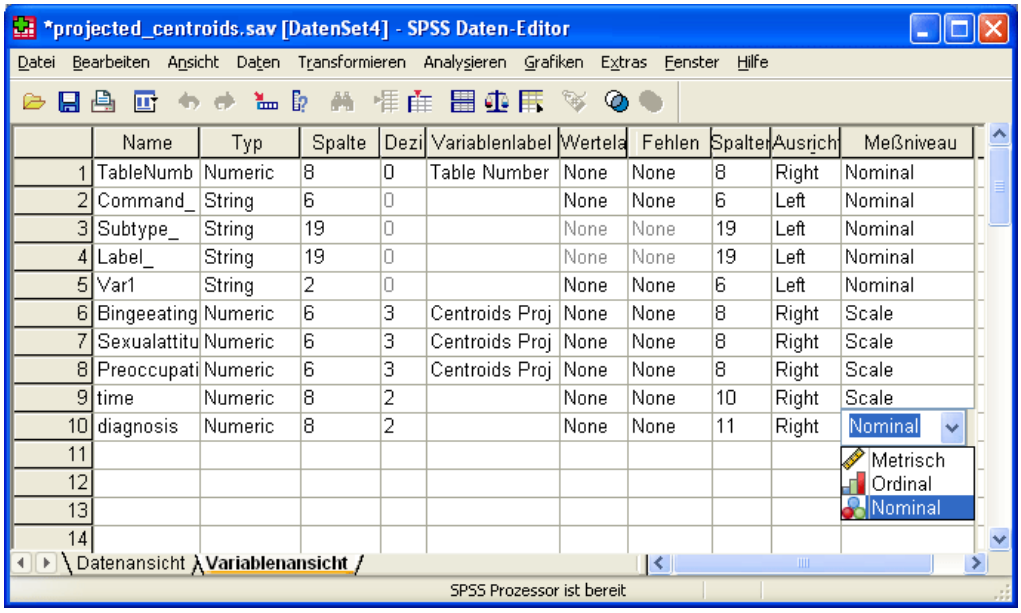
- Geben Sie *time* als Zielvariable ein.
- Geben Sie trunc(($casenum-1)/4) + 1 als numerischen Ausdruck ein.
- Klicken Sie auf OK.

Abbildung 10-47
Dialogfeld "Variable berechnen"



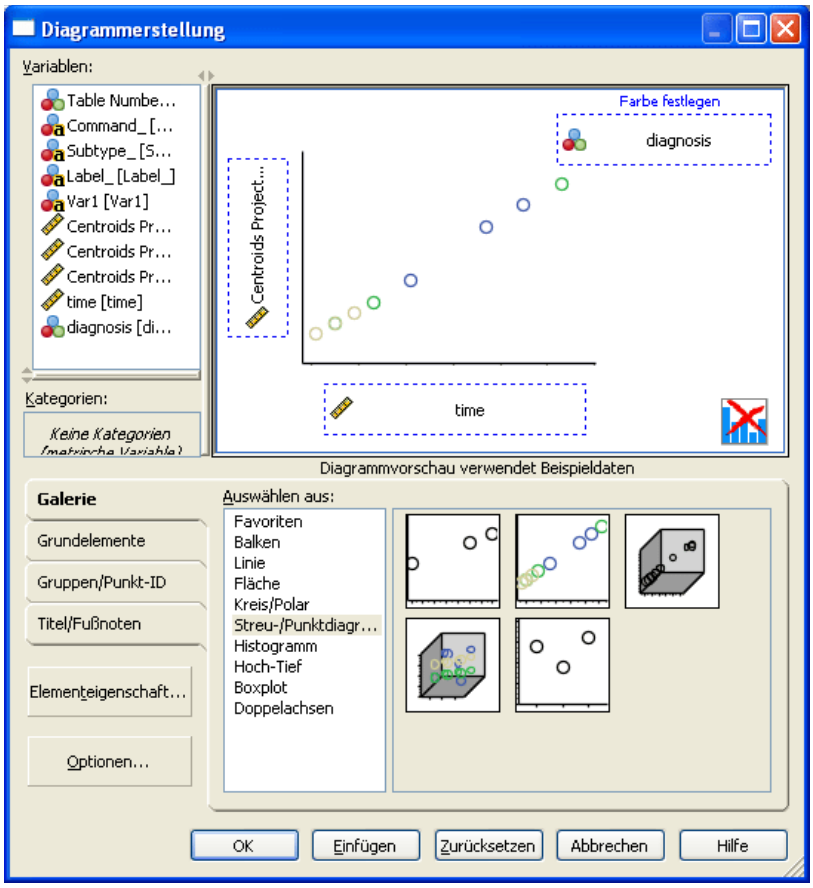
- Rufen Sie das Dialogfeld "Variable berechnen" auf.
- Geben Sie *diagnosis* als Zielvariable ein.
- Geben Sie mod($casenum-1, 4) + 1 als numerischen Ausdruck ein.
- Klicken Sie auf OK.

Abbildung 10-48
Projected_centroids.sav


*projected_centroids.sav [DatenSet4] - SPSS Daten-Editor
Datei Bearbeiten Ansicht Daten Transformieren Analysieren Grafiken Extras Fenster Hilfe
Name Typ Spalte Dezi Variablenlabel Wertela Fehlen Spalter Ausrich Meßniveau
1 TableNumb Numeric 8 0 Table Number None None 8 Right Nominal
2 Command_ String 6 0 None None 6 Left Nominal
3 Subtype_ String 19 0 None None 19 Left Nominal
4 Label_ String 19 0 None None 19 Left Nominal
5 Var1 String 2 0 None None 6 Left Nominal
6 Bingeeating Numeric 6 3 Centroids Proj None None 8 Right Scale
7 Sexualattitu Numeric 6 3 Centroids Proj None None 8 Right Scale
8 Preoccupati Numeric 6 3 Centroids Proj None None 8 Right Scale
9 time Numeric 8 2 None None 10 Right Scale
10 diagnosis Numeric 8 2 None None 11 Right Nominal
11
12
13
14
Metrisch
Ordinal
Nominal
Datenansicht Variablenansicht
SPSS Prozessor ist bereit

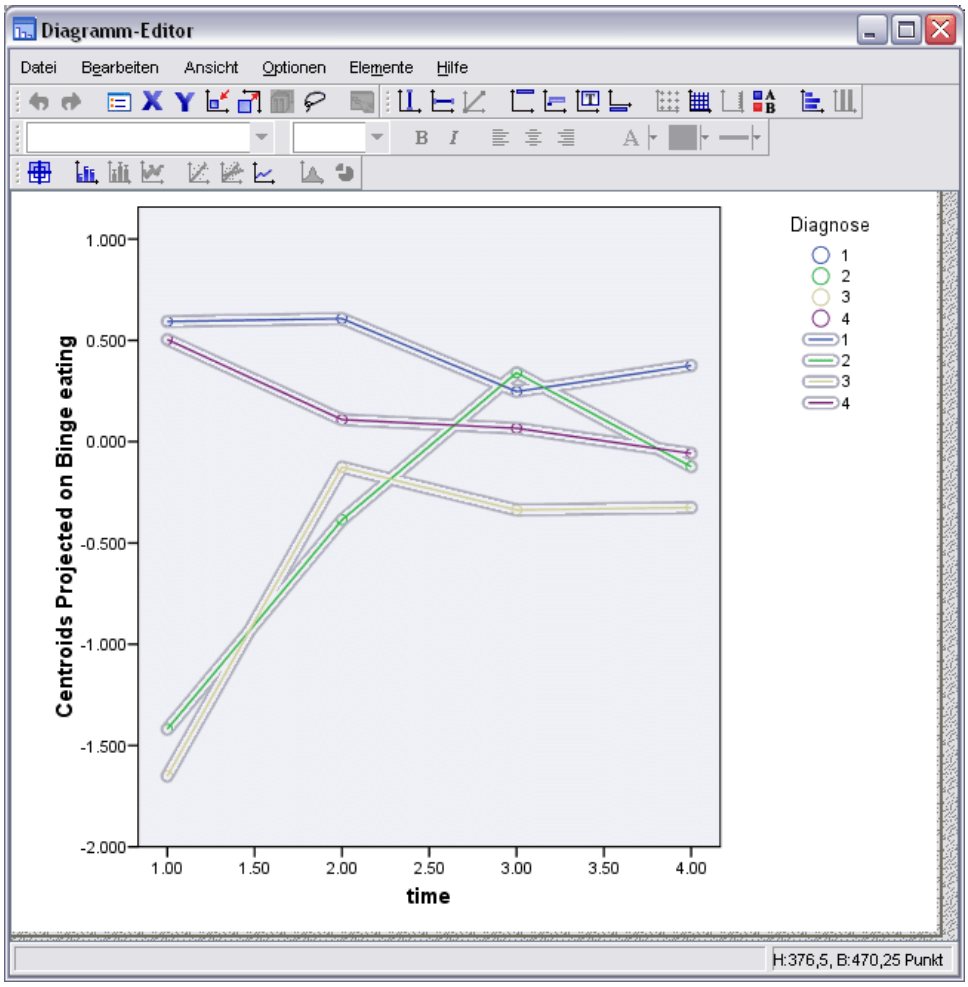
Ändern Sie in der Variablenansicht das Maß für *diagnosis* (Diagnose) von “Metrisch” in “Nominal”.

Abbildung 10-49
Diagrammerstellung



- Um schließlich die auf die “Binge eating” (Fressattacken) projizierten Zentroide des Diagnosezeitpunkts im Verlauf der Zeit zu betrachten, müssen Sie die Diagrammerstellung erneut öffnen und durch Klicken auf Zurücksetzen die vorherige Auswahl löschen.
- Wählen Sie die Galerie Streu-/Punktdiagramm aus und wählen Sie die Option “Gruppiertes Streudiagramm”.
- Wählen Sie *Centroids Projected on Binge eating* (Zentroide auf Freßattacken projiziert) als *y*-Achsen-Variable und *time* (Zeit) als *x*-Achsen-Variable aus.
- Wählen Sie aus, dass die Farben anhand von *diagnosis* (Diagnose) festgelegt werden sollen.
- Klicken Sie auf OK.

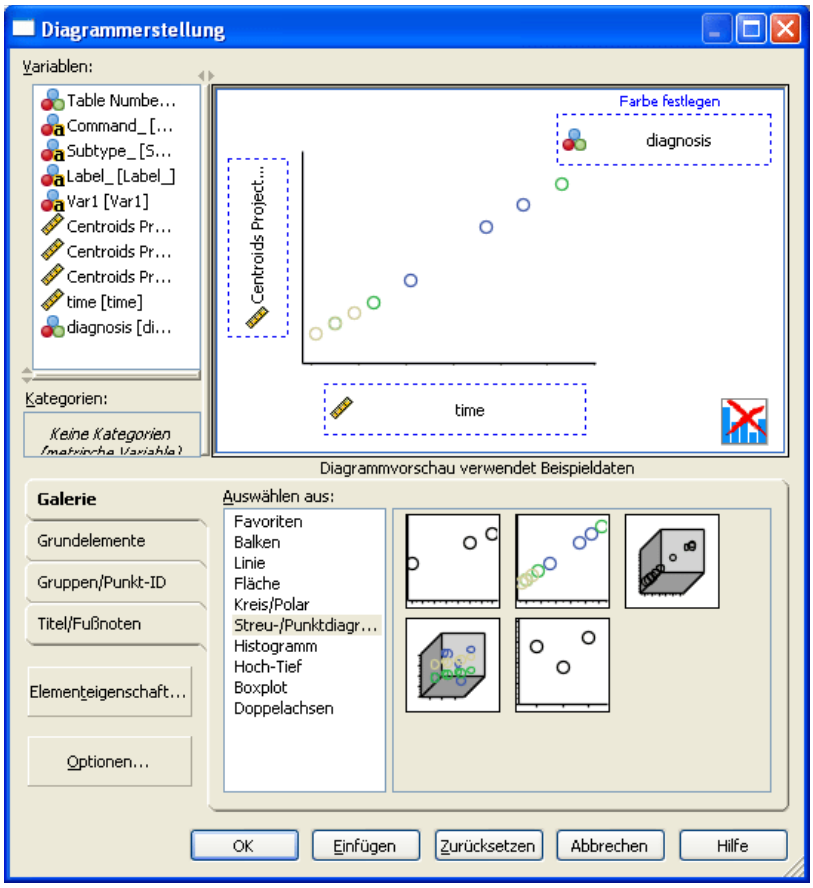
Abbildung 10-50
Auf "Binge eating" (Freßattacken) projizierte Zentroide des Diagnosezeitpunkts im Verlauf der Zeit



- Doppelklicken Sie dann zur Verbindung der Punkte auf das Diagramm und klicken Sie im Diagramm-Editor auf das Tool zum Hinzufügen von Interpolationslinien.
- Schließen Sie den Diagramm-Editor.

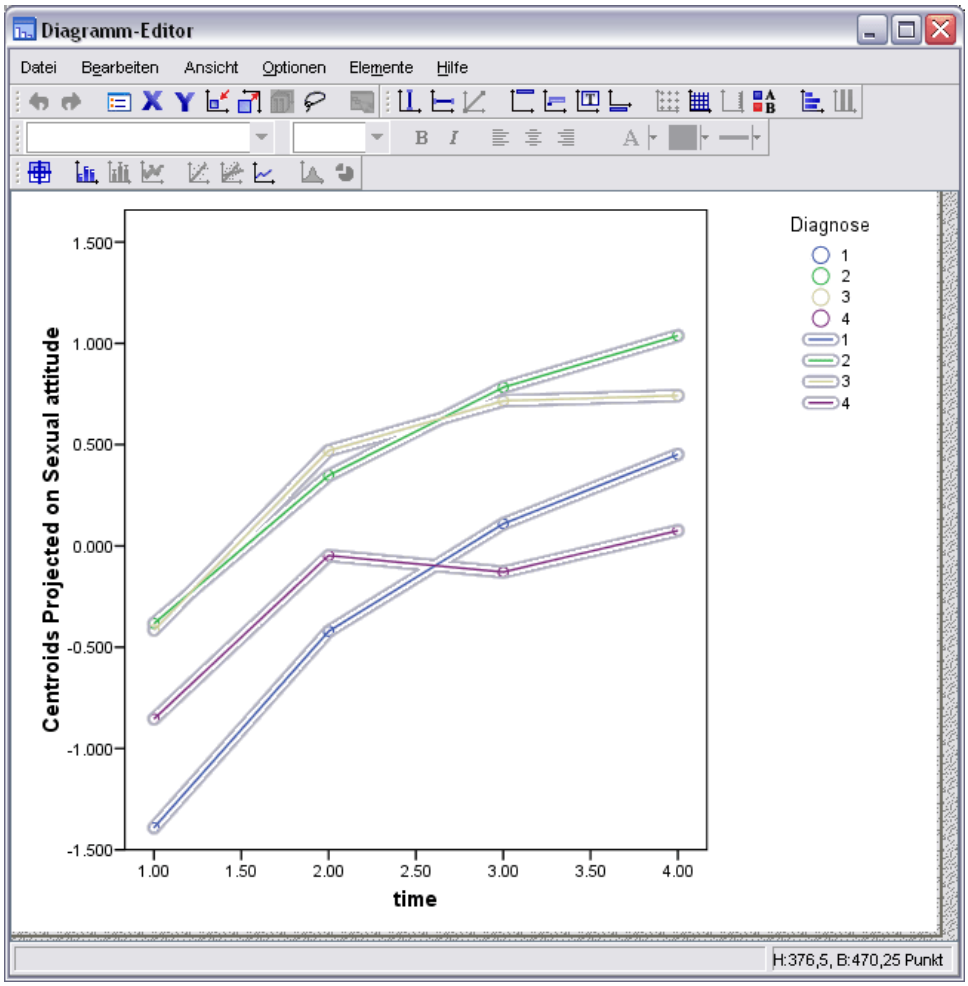
Was die Freßattacken betrifft, so ist klar, dass die anorektischen Gruppen andere Anfangswerte aufweisen als die bulimischen Gruppen. Dieser Unterschied nimmt im Laufe der Zeit ab, da sich die anorektischen Gruppen kaum ändern, während bei den bulimischen Gruppen ein Fortschritt zu verzeichnen ist.

Abbildung 10-51
Diagrammerstellung



- Rufen Sie die Diagrammerstellung erneut auf.
- Heben Sie die Auswahl von *Centroids Projected on Binge eating* (Zentroide auf Freßattacken projiziert) als *y*-Achsen-Variable auf und wählen Sie statt dessen *Centroids Projected on Sexual attitude* (Zentroide auf sexuelle Einstellung projiziert) als *y*-Achsen-Variable aus.
- Klicken Sie auf OK.

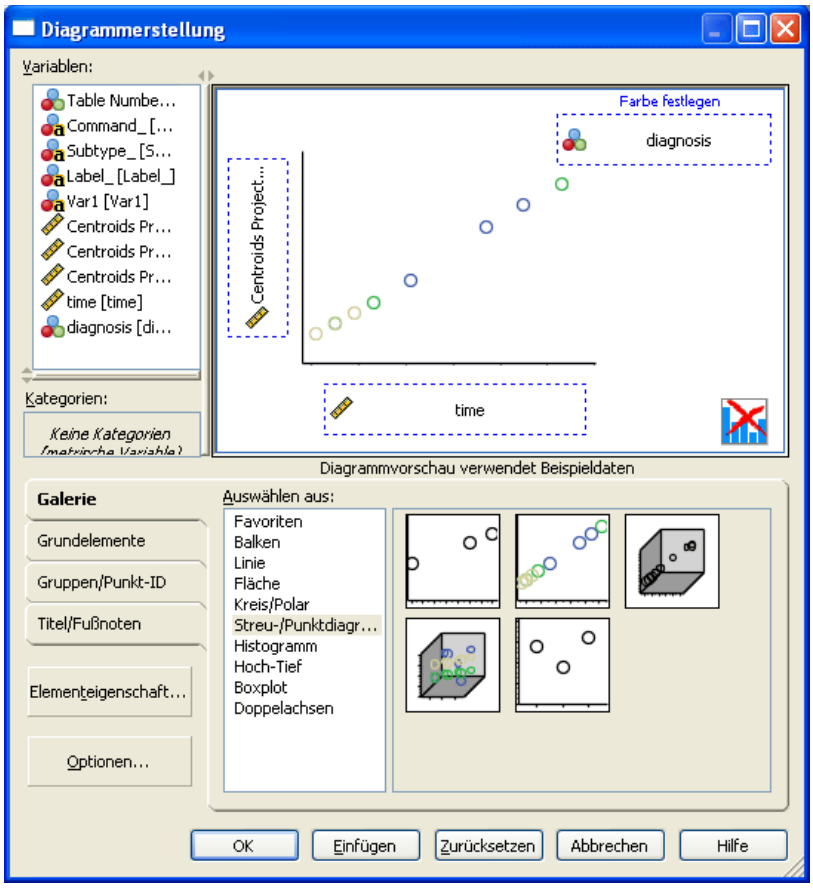
Abbildung 10-52
Auf "Sexual attitude" (sexuelle Einstellung) projizierte Zentroide des Diagnosezeitpunkts im Verlauf der Zeit



- Doppelklicken Sie dann zur Verbindung der Punkte auf das Diagramm und klicken Sie im Diagramm-Editor auf das Tool zum Hinzufügen von Interpolationslinien.
- Schließen Sie den Diagramm-Editor.

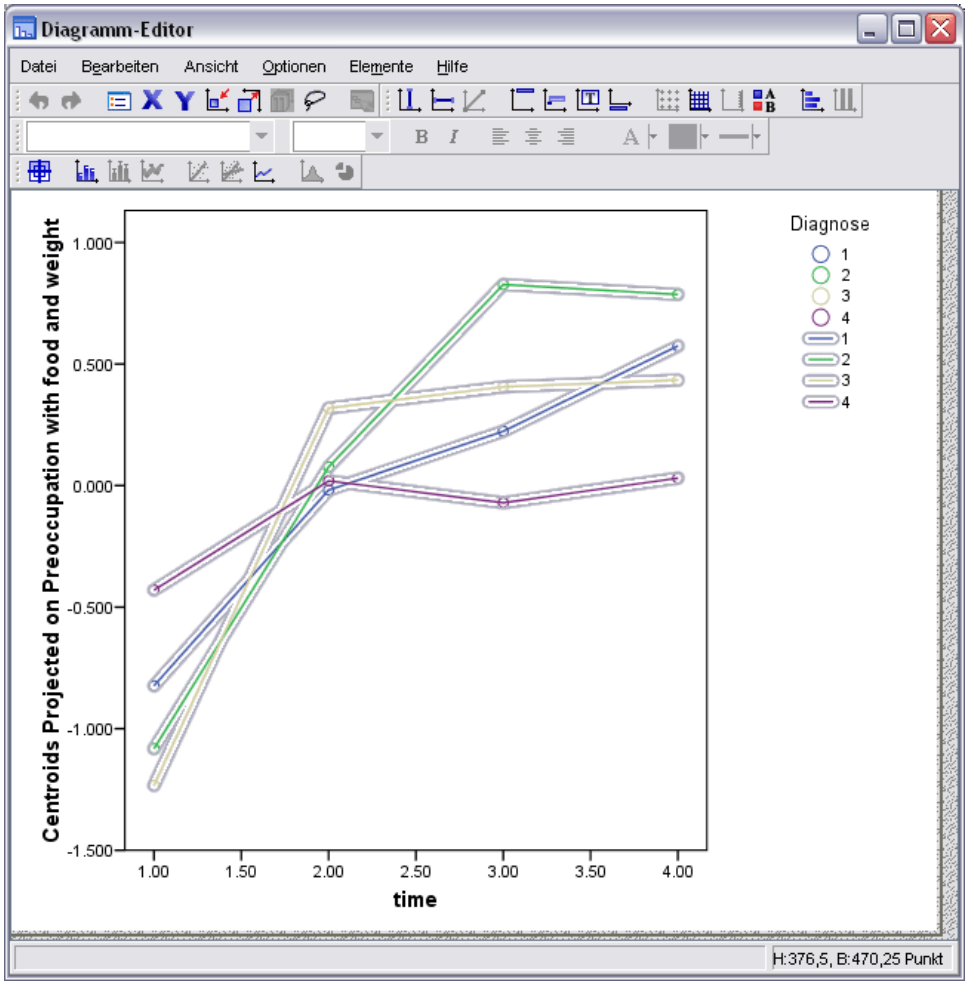
Bei der sexuellen Einstellungen verlaufen die vier Kurven mehr oder weniger parallel und alle Gruppen zeigen Fortschritte. Die bulimischen Gruppen weisen allerdings höhere (bessere) Werte auf als die anorektische Gruppe.

Abbildung 10-53
Diagrammerstellung



- Rufen Sie die Diagrammerstellung erneut auf.
- Heben Sie die Auswahl von *Centroids Projected on Sexual attitude* (Zentroide auf sexuelle Einstellung projiziert) als *y*-Achsen-Variable auf und wählen Sie statt dessen *Centroids Projected on Preoccupation with food and weight* (Zentroide auf Beschäftigung mit Essen und Gewicht projiziert) als *y*-Achsen-Variable aus.
- Klicken Sie auf OK.

Abbildung 10-54
Auf "Preoccupation with food and weight" (Beschäftigung mit Essen und Gewicht) projizierte Zentroide des Diagnosezeitpunkts im Verlauf der Zeit



- Doppelklicken Sie dann zur Verbindung der Punkte auf das Diagramm und klicken Sie im Diagramm-Editor auf das Tool zum Hinzufügen von Interpolationslinien.
- Schließen Sie den Diagramm-Editor.

"Preoccupation with food and weight" (Beschäftigung mit Essen und Gewicht) ist eine Variable, die die Kernsymptome repräsentiert, die allen vier Gruppen gemeinsam sind. Anders als bei den Patienten mit atypischer Eßstörung weisen die anorektische Gruppe und die beiden bulimischen Gruppen sowohl am Anfang als auch am Ende sehr ähnliche Niveaus auf.

Empfohlene Literatur

In folgenden Texten finden Sie weitere Informationen zur kategorialen Hauptkomponentenanalyse:

De Haas, M., J. A. Algera, H. F. J. M. Van Tuijl, als auch J. J. Meulman. 2000. Macro and micro goal setting: In search of coherence. *Applied Psychology*, 49, 579–595.

De Leeuw, J. 1982. Nonlinear principal components analysis. In: *COMPSTAT Proceedings in Computational Statistics,* Wien: Physica Verlag, 77–89.

Eckart, C., als auch G. Young. 1936. The approximation of one matrix by another one of lower rank. *Psychometrika*, 1, 211–218.

Gabriel, K. R. 1971. The biplot graphic display of matrices with application to principal components analysis. *Biometrika*, 58, 453–467.

Gifi, A. 1985. *PRINCALS. Research Report UG-85-02*. Leiden: Department of Data Theory, Universität Leiden.

Gower, J. C., als auch J. J. Meulman. 1993. The treatment of categorical information in physical anthropology. *International Journal of Anthropology*, 8, 43–51.

Heiser, W. J., als auch J. J. Meulman. 1994. Homogeneity analysis: Exploring the distribution of variables and their nonlinear relationships. In: *Correspondence Analysis in the Social Sciences: Recent Developments and Applications,* M. Greenacre, als auch J. Blasius (Hgg.). New York: Academic Press, 179–209.

Kruskal, J. B. 1978. Factor analysis and principal components analysis: Bilinear methods. In: *International Encyclopedia of Statistics,* W. H. Kruskal, als auch J. M. Tanur (Hgg.). New York: The Free Press, 307–330.

Kruskal, J. B., als auch R. N. Shepard. 1974. A nonmetric variety of linear factor analysis. *Psychometrika*, 39, 123–157.

Meulman, J. J. 1993. Principal coordinates analysis with optimal transformations of the variables: Minimizing the sum of squares of the smallest eigenvalues. *British Journal of Mathematical and Statistical Psychology*, 46, 287–300.

Meulman, J. J., als auch P. Verboon. 1993. Points of view analysis revisited: Fitting multidimensional structures to optimal distance components with cluster restrictions on the variables. *Psychometrika*, 58, 7–35.

Meulman, J. J., A. J. Van der Kooij, als auch A. Babinec. 2000. New features of categorical principal components analysis for complicated data sets, including data mining. In: *Classification, Automation and New Media,* W. Gaul, als auch G. Ritter (Hgg.). Berlin: Springer-Verlag, 207–217.

Meulman, J. J., A. J. Van der Kooij, als auch W. J. Heiser. 2004. Principal components analysis with nonlinear optimal scaling transformations for ordinal and nominal data. In: *Handbook of Quantitative Methodology for the Social Sciences,* D. Kaplan (Hg.). Thousand Oaks, Kalifornien: Sage Publications, Inc, 49–70.

Theunissen, N. C. M., J. J. Meulman, A. L. Den Ouden, H. M. Koopman, G. H. Verrips, S. P. Verloove-Vanhorick, als auch J. M. Wit. 2003. Changes can be studied when the measurement instrument is different at different time points. *Health Services and Outcomes Research Methodology*, 4, 109–126.

Tucker, L. R. 1960. Intra-individual and inter-individual multidimensionality. In: *Psychological Scaling: Theory & Applications,* H. Gulliksen, als auch S. Messick (Hgg.). New York: John Wiley and Sons, 155–167.

Vlek, C., als auch P. J. Stallen. 1981. Judging risks and benefits in the small and in the large. *Organizational Behavior and Human Performance*, 28, 235–271.

Wagenaar, W. A. 1988. *Paradoxes of gambling behaviour*. London: Lawrence Erlbaum Associates, Inc.

Young, F. W., Y. Takane, als auch J. De Leeuw. 1978. The principal components of mixed measurement level multivariate data: An alternating least squares method with optimal scaling features. *Psychometrika*, 43, 279–281.

Zeijl, E., Y. te Poel, M. du Bois-Reymond, J. Ravesloot, als auch J. J. Meulman. 2000. The role of parents and peers in the leisure activities of young adolescents. *Journal of Leisure Research*, 32, 281–302.

Kapitel
11

Nichtlineare kanonische Korrelationsanalyse

Die nichtlineare kanonische Korrelationsanalyse dient zur Bestimmung, wie ähnlich zwei oder mehr Gruppen von Variablen einander sind. Wie bei der linearen kanonischen Korrelationsanalyse ist das Ziel, die Varianz in den Beziehungen unter den Gruppen in einem flachdimensionierten Raum so gut wie möglich zu berücksichtigen. Anders als bei dei der linearen kanonischen Korrelationsanalyse wird bei der nichtlinearen kanonischen Korrelationsanalyse nicht von einem linearen Messniveau oder von linearen Beziehungen ausgegangen. Ein weiterer wichtiger Unterschied besteht darin, dass bei der nichtlinearen kanonischen Korrelationsanalyse die Ähnlichkeit zwischen den Gruppen durch den gleichzeitigen Vergleich linearer Kombinationen der Variablen in jeder Gruppe mit einer unbekannten Gruppe durchgeführt wird (Objektwerte).

Beispiel: Analyse der Umfrageergebnisse

Das Beispiel in diesem Kapitel stammt aus einer Umfrage (Verdegaal, 1985). Die Antworten von 15 Subjekten auf 8 Variablen wurden aufgezeichnet. Die Variablen, Variablenlabels und Wertelabels (Kategorien) im Daten-Set werden in der folgenden Tabelle angezeigt.

Tabelle 11-1
Umfragedaten

Variablenname	Variablenlabel	Wertelabels
age	*Age in years*	20–25, 26–30, 31–35, 36–40, 41–45, 46–50, 51–55, 56–60, 61–65, 66–70
marital	*Marital status*	Single, Married, Other
pet	*Pets owned*	No, Cat(s), Dog(s), Other than cat or dog, Various domestic animals
news	*Newspaper read most often*	None, Telegraaf, Volkskrant, NRC, Other
music	*Music preferred*	Classical, New wave, Popular, Variety, Don't like music
live	*Neighborhood preference*	Town, Village, Countryside
math	*Math test score*	0–5, 6–10, 11–15
language	*Language test score*	0–5, 6–10, 11–15, 16–20

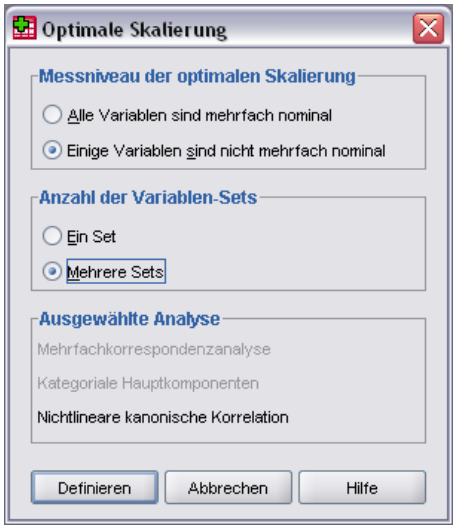
Dieses Daten-Set finden Sie in *verd1985.sav*. Für weitere Informationen siehe Beispieldateien in Anhang A auf S. 326. Die ersten sechs Variablen sind relevant; diese Variablen sind in drei Sets gegliedert. Set 1 umfasst *alter* und *heirat*, Set 2 besteht aus *pet* und *news* und in Set 3 finden sich *music* und *live*. Die Variable *pet* ist mehrfach nominal skaliert, die Variable *age* ordinal. Alle anderen Variablen sind einzeln nominal skaliert. Für diese Analyse ist eine zufällige Anfangskonfiguration erforderlich. Standardmäßig ist diese Anfangskonfiguration numerisch. Wird ein Teil der Variablen jedoch als einzeln nominal behandelt, ohne dass eine Möglichkeit zum Sortieren besteht, sollten Sie eine zufällige Anfangskonfiguration verwenden. Bei den meisten Variablen in dieser Untersuchung ist dies der Fall.

Untersuchen der Daten

- Zum Erstellen einer nichtlinearen kanonischen Korrelationsanalyse für dieses Daten-Set wählen Sie die folgenden Befehle aus den Menüs aus:

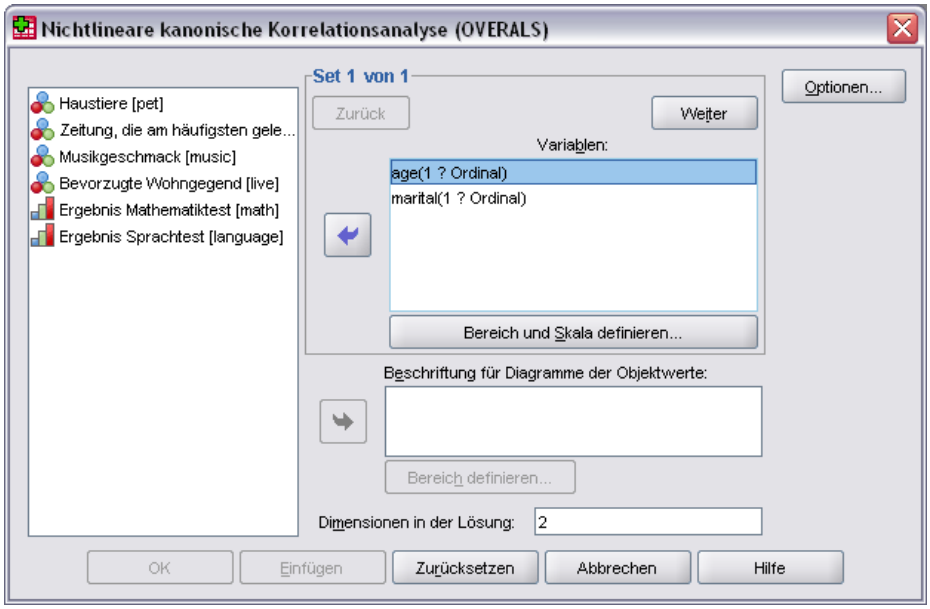
  Analysieren
  Dimensionsreduktion
  Optimale Skalierung...

Abbildung 11-1
Dialogfeld "Optimale Skalierung"



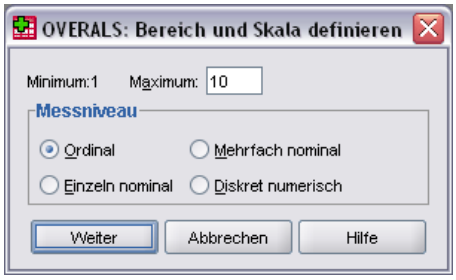
- Wählen Sie im Gruppenfeld "Messniveau der optimalen Skalierung" die Option Einige Variablen sind nicht mehrfach nominal.
- Wählen Sie in der Gruppe "Anzahl der Variablen-Sets" die Option Mehrere Sets.
- Klicken Sie auf Definieren.

Abbildung 11-2
Dialogfeld "Nichtlineare kanonische Korrelationsanalyse"



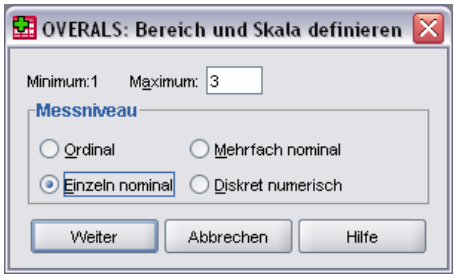
- Wählen Sie *Age in years* und *Marital status* als Variablen für das erste Set.
- Wählen Sie *age* aus und klicken Sie auf Bereich und Skala definieren.

Abbildung 11-3
Dialogfeld "Bereich und Skala definieren"



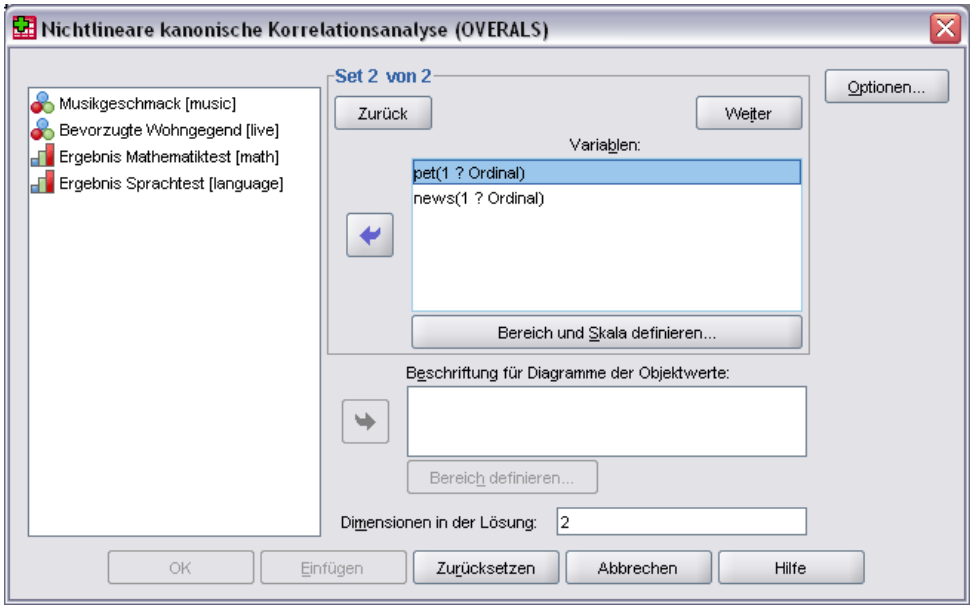
- Geben Sie 10 als Höchstwert für diese Variable ein.
- Klicken Sie auf Weiter.
- Wählen Sie im Dialogfeld "Nichtlineare kanonische Korrelationsanalyse" die Variable *marital* aus und klicken Sie auf Bereich und Skala definieren.

Abbildung 11-4
Dialogfeld "Bereich und Skala definieren"



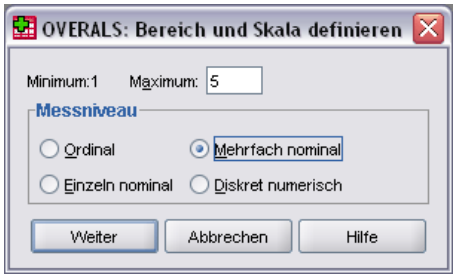
- Geben Sie 3 als Höchstwert für diese Variable ein.
- Wählen Sie Einzeln nominal als Messniveau aus.
- Klicken Sie auf Weiter.
- Klicken Sie im Dialogfeld "Nichtlineare kanonische Korrelationsanalyse" auf Weiter, um das nächste Variablen-Set zu definieren.

Abbildung 11-5
Dialogfeld "Nichtlineare kanonische Korrelationsanalyse"



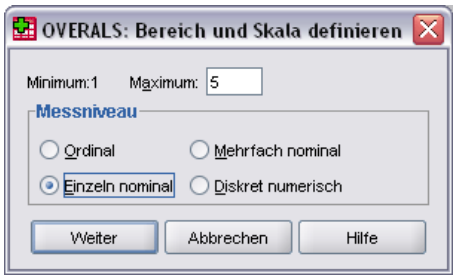
- Wählen Sie *Pets owned* und *Newspaper read most often* als Variablen für das zweite Set aus.
- Wählen Sie *pet* aus und klicken Sie auf Bereich und Skala definieren.

Abbildung 11-6
Dialogfeld "Bereich und Skala definieren"



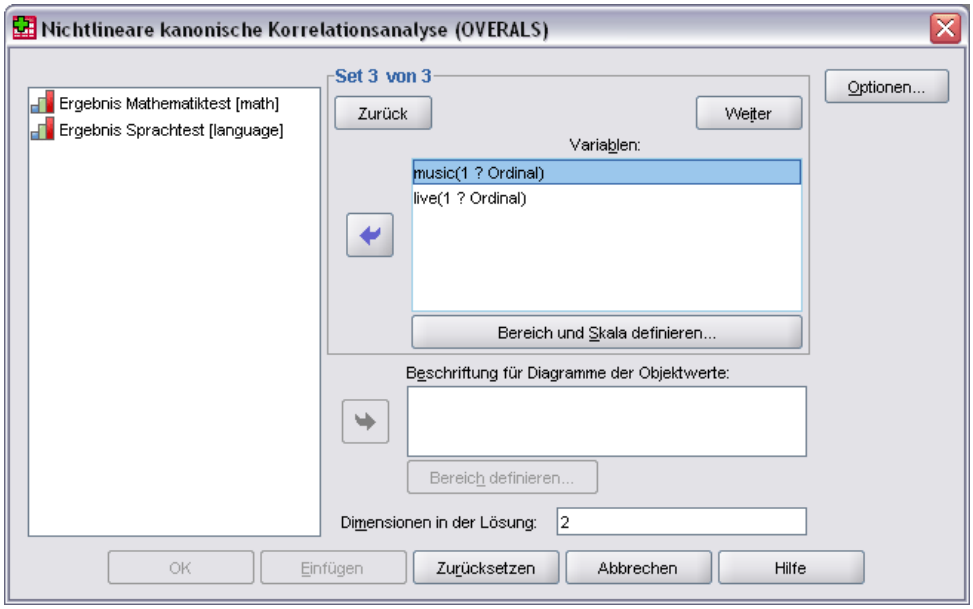
- Geben Sie 5 als Höchstwert für diese Variable ein.
- Wählen Sie Mehrfach nominal als Messniveau aus.
- Klicken Sie auf Weiter.
- Wählen Sie im Dialogfeld "Nichtlineare kanonische Korrelationsanalyse" die Variable *news* aus und klicken Sie auf Bereich und Skala definieren.

Abbildung 11-7
Dialogfeld "Bereich und Skala definieren"



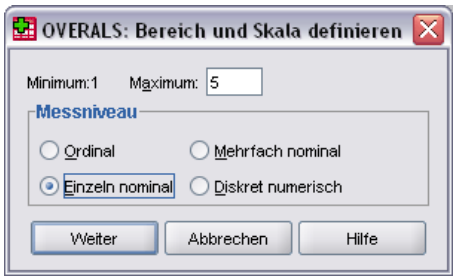
- Geben Sie 5 als Höchstwert für diese Variable ein.
- Wählen Sie Einzeln nominal als Messniveau aus.
- Klicken Sie auf Weiter.
- Klicken Sie im Dialogfeld "Nichtlineare kanonische Korrelationsanalyse" auf Weiter, um das letzte Variablen-Set zu definieren.

Abbildung 11-8
Dialogfeld "Nichtlineare kanonische Korrelationsanalyse"



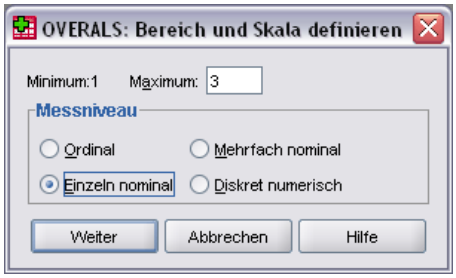
- Wählen Sie *Music preferred* und *Neighborhood preference* als Variablen für das dritte Set aus.
- Wählen Sie *music* aus und klicken Sie auf Bereich und Skala definieren.

Abbildung 11-9
Dialogfeld "Bereich und Skala definieren"



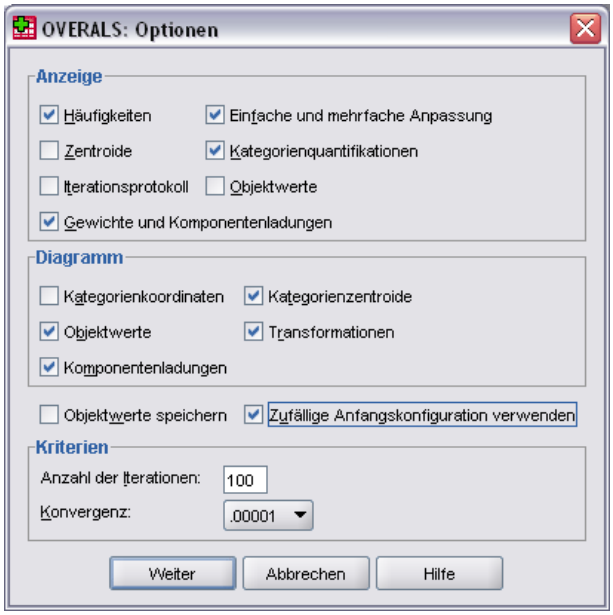
- Geben Sie 5 als Höchstwert für diese Variable ein.
- Wählen Sie Einzeln nominal als Messniveau aus.
- Klicken Sie auf Weiter.
- Wählen Sie im Dialogfeld "Nichtlineare kanonische Korrelationsanalyse" die Variable *live* aus und klicken Sie auf Bereich und Skala definieren.

Abbildung 11-10
Dialogfeld "Bereich und Skala definieren"



- Geben Sie 3 als Höchstwert für diese Variable ein.
- Wählen Sie Einzeln nominal als Messniveau aus.
- Klicken Sie auf Weiter.
- Klicken Sie im Dialogfeld "Nichtlineare kanonische Korrelationsanalyse" auf Optionen.

Abbildung 11-11
Dialogfeld "Optionen"



- Deaktivieren Sie die Option Zentroide und wählen Sie in der Gruppe "Anzeige" die Option Gewichte und Komponentenladungen.
- Wählen Sie in der Gruppe "Diagramm" die Optionen Kategorienzentroide und Transformationen.
- Wählen Sie Zufällige Anfangskonfiguration verwenden.
- Klicken Sie auf Weiter.
- Klicken Sie im Dialogfeld "Nichtlineare kanonische Korrelationsanalyse" auf OK.

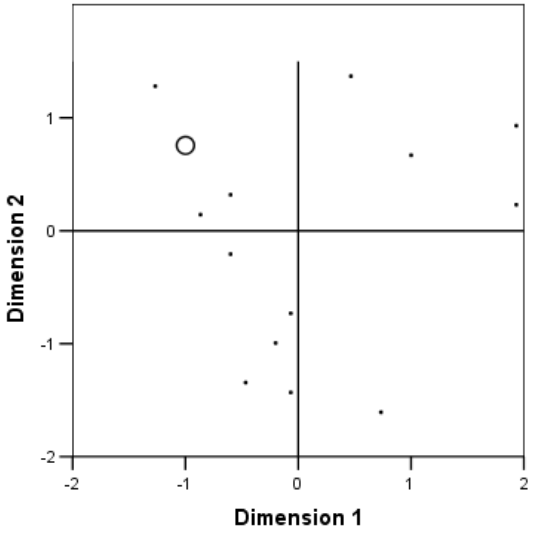
Bei der kategorialen kanonischen Korrelationsanalyse mit optimaler Skalierung wird zunächst eine Liste der Variablen erzeugt, in der das zugehörige Messniveau mit optimaler Skalierung vermerkt ist, und dann eine Tabelle, aus der die Häufigkeit der Objekte in den Kategorien hervorgeht. Diese Tabelle ist besonders wichtig, wenn Daten fehlen, weil nahezu leere Kategorien die Lösung eher dominieren. In diesem Beispiel fehlen keine Daten.

Bei einer zweiten Vorabüberprüfung untersuchen Sie, ob Ausreißer im Diagramm der Objektwerte vorliegen. Die Quantifikationen von Ausreißern weichen so deutlich von anderen Objekten ab, dass diese Elemente an den Grenzen des Diagramms liegen und so eine oder auch mehrere Dimensionen dominieren.

Bei Ausreißern stehen zwei Möglichkeiten zur Auswahl. Sie können die Ausreißer schlicht aus den Daten entfernen und dann die nichtlineare kanonische Korrelationsanalyse wiederholen. Sie können die extremen Antworten der Ausreißerobjekte aber auch umkodieren, indem Sie einige Kategorien zusammenfassen (zusammenführen).

Aus dem Diagramm der Objektwerte geht hervor, dass keine Ausreißer für die Untersuchungsdaten vorhanden sind.

Abbildung 11-12
Objektwerte



Fälle gewichtet nach Anzahl der Objekte.

Berücksichtigen der Ähnlichkeiten zwischen Sets

Der Zusammenhang zwischen den Sets in einer nichtlinearen kanonischen Korrelationsanalyse kann auf verschiedene Weise gemessen werden; die Ergebnisse werden dabei jeweils in einer separaten Tabelle oder Gruppe von Tabellen aufgeführt.

Zusammenfassung der Analyse

Die Anpassungs- und Verlustwerte zeigen die Anpassungsgüte der nichtlinearen kanonischen Korrelationsanalyselösung für die optimal quantifizierten Daten im Hinblick auf die Zusammenhänge zwischen den Sets. Die Tabelle mit der Zusammenfassung der Analyse enthält den Anpassungswert, die Verlustwerte und die Eigenwerte für die Beispieluntersuchung.

Abbildung 11-13
Zusammenfassung der Analyse

		Dimension		
		1	2	Summe
Verlustfunktion	Set 1	,240	,183	,423
	Set 2	,184	,408	,593
	Set 3	,171	,205	,376
	Mittelwert	,199	,265	,464
Eigenwert		,801	,735	
Anpassung				1,536

Der Verlust wird auf die Dimensionen und Sets aufgeteilt. Bei jeder Dimension und jedem Set bezeichnet der Verlust den Anteil an der Variation in den Objektwerten, der nicht durch die gewichtete Kombination der Variablen im Set abgedeckt ist. Der durchschnittliche Verlust wird als Mittelwert gekennzeichnet. In diesem Beispiel liegt der durchschnittliche Verlust über die Sätze bei 0,464. Beachten Sie, dass in der zweiten Dimension höhere Verluste auftreten als in der ersten Dimension.

Der Eigenwert für jede Dimension ist gleich 1 minus dem durchschnittlichen Verlust für die Dimension. Dieser Wert gibt den Teil der Beziehung an, der durch die einzelnen Dimensionen dargestellt wird. Die Eigenwerte ergeben zusammengerechnet die Gesamtanpassung. Bei den Daten nach Verdegaal werden 0,801 / 1,536 = 52 % der tatsächlichen Anpassung durch die erste Dimension erzielt.

Der maximale Anpassungswert ist gleich der Anzahl der Dimensionen. Wird dieser Maximalwert erreicht, bedeutet dies, dass die Beziehung perfekt ist. Der durchschnittliche Verlustwert über die Sätze und Dimensionen zeigt die Differenz zwischen der maximalen Anpassung und der tatsächlichen Anpassung. Die Summe aus Anpassung und durchschnittlichem Verlust ist gleich der Anzahl der Dimensionen. Eine perfekte Ähnlichkeit liegt nur selten vor und konzentriert sich in der Regel nur auf triviale Aspekte in den Daten.

Die kanonische Korrelation ist eine weitere, beliebte Statistik mit zwei Variablen-Sets. Die kanonische Korrelation hängt mit dem Eigenwert zusammen und liefert somit keine zusätzlichen Informationen; daher wird diese Korrelation nicht in der Ausgabe der nichtlinearen kanonischen Korrelationsanalyse aufgeführt Bei zwei Variablen-Sets wird die kanonische Korrelation pro Dimension mit der folgenden Formel berechnet:

$$\rho_d = 2 \times E_d - 1$$

Hierbei gilt: *d* ist die Anzahl der Dimensionen und *E* ist der Eigenwert.

Mit der folgenden Formel können Sie die kanonische Korrelation für mehr als zwei Sets verallgemeinern:

$$\rho_d = ((K \times E_d) - 1)/(K - 1)$$

Hierbei gilt: *d* ist die Anzahl der Dimensionen, *K* ist die Anzahl der Sets und *E* ist der Eigenwert. Im obigen Beispiel:

$$\rho_1 = ((3 \times 0.801) - 1)/2 = 0.702$$

und

$$\rho_2 = ((3 \times 0.735) - 1)/2 = 0.603$$

Gewichte und Komponentenladungen

Als weiteres Zusammenhangsmaß steht außerdem die Mehrfachkorrelation zwischen linearen Kombinationen aus jedem Set und den Objektwerten zur Verfügung. Wenn ein Set keine mehrfach nominalen Variablen enthält, können Sie dieses Maß folgendermaßen berechnen: Multiplizieren Sie die Gewichtung und die Komponentenladung der einzelnen Variablen im Set, addieren Sie diese Produkte und ziehen Sie dann die Quadratwurzel aus der Summe.

Abbildung 11-14
Gewichtungen

Set		Dimension 1	Dimension 2
1	Age in years	,680	,789
	Marital status	,296	-1,016
2	Newspaper read most often	-,845	-,361
3	Music preferred	,631	-,749
	Neighborhood preference	-,484	-,780

Abbildung 11-15
Komponentenladungen

Set			Dimension 1	Dimension 2
1	Age in years[a,b]		,834	,259
	Marital status[c,b]		,651	-,604
2	Pets owned[d,e]	Dimension 1	,397	-,431
		2	-,277	,680
	Newspaper read most often[c,b]		-,667	-,391
3	Music preferred[c,b]		,786	-,500
	Neighborhood preference[c,b]		-,687	-,540

a. Optimales Skalierungsniveau: Ordinal

b. Projektionen der einfach quantifizierten Variablen im Objektraum

c. Optimales Skalierungsniveau: Einfach nominal

d. Optimales Skalierungsniveau: mehrfach nominal

e. Projektionen der mehrfach quantifizierten Variablen im Objektraum

Diese Zahlen geben die Gewichtungen und die Komponentenladungen für die Variablen in diesem Beispiel an. Die Mehrfachkorrelation (*R*) für die erste gewichtete Summe optimal skalierter Variablen (*Age in years* und *Marital status*) in der ersten Dimension der Objekte lautet:

$$\begin{aligned} R &= \sqrt{(0.701 \times 0.841 + (-0.273 \times -0.631))} \\ &= \sqrt{(0.5895 + 0.1723)} \\ &= 0.873 \end{aligned}$$

Für jede Dimension gilt: 1 – Verlust = R^2. Beispiel: Aus der Tabelle mit der Zusammenfassung der Analyse ergibt sich: 1 – 0,238 = 0,762, also 0,873 im Quadrat (plus Rundungsfehler). Kleine Verlustwerte weisen somit auf umfangreiche Mehrfachkorrelationen zwischen den gewichteten Summen optimal skalierter Variablen und Dimensionen hin. Die Gewichtungen sind bei mehrfach nominalen Variablen nicht eindeutig. Bei mehrfach nominalen Variablen gilt die Berechnung: 1 – Verlust pro Set.

Aufteilen der Anpassung und der Verluste

Der Verlust in den einzelnen Sets wird bei der nichtlinearen kanonischen Korrelationsanalyse auf verschiedene Weise aufgeteilt. Die Anpassungstabelle enthält die Tabellen für Mehrfachanpassung, Einzelanpassung und Einzelverlust, die durch die nichtlineare kanonische Korrelationsanalyse für die Beispieluntersuchung erzeugt wurden. Beachten Sie: Die Mehrfachanpassung minus die Einzelanpassung ergibt den Einzelverlust.

Abbildung 11-16
Aufteilen der Anpassung und der Verluste

		Mehrfachanpassung			Einzelne Anpassung			Einzelne Verlustfunktion		
		Dimension			Dimension			Dimension		
Set		1	2	Summe	1	2	Summe	1	2	Summe
1	Age in years[a]	,494	,676	1,170	,462	,622	1,085	,032	,054	,085
	Marital status[b]	,089	1,033	1,122	,088	1,033	1,120	,001	,000	,001
2	Pets owned[c]	,402	,439	,841						
	Newspaper read most often[b]	,724	,187	,911	,714	,130	,844	,010	,057	,067
3	Music preferred[b]	,421	,577	,998	,398	,561	,960	,022	,016	,039
	Neighborhood preference[b]	,234	,609	,843	,234	,608	,843	,000	,000	,000

a. Optimales Skalierungsniveau: Ordinal

b. Optimales Skalierungsniveau: Einfach nominal

c. Optimales Skalierungsniveau: mehrfach nominal

Der Einzelverlust zeigt den Verlust, der sich aus dem Einschränken der Variablen auf eine einzige Quantifikationsgruppe ergibt (also einzeln nominal, ordinal oder numerisch). Liegt ein hoher Einzelverlust vor, sollten die Variablen eher als mehrfach nominal behandelt werden. In diesem Beispiel sind die Einzelanpassung und die Mehrfachanpassung jedoch nahezu identisch; die Mehrfachkoordinaten liegen also nahezu auf einer geraden Linie in der Richtung, die durch die Gewichtungen vorgegeben ist.

Die Mehrfachanpassung ist gleich der Varianz der Mehrfachkategorienkoordinaten für die einzelnen Variablen. Diese Maßzahlen entsprechen den Diskriminationsmaßen bei der Homogenitätsanalyse. Aus der Tabelle für die Mehrfachanpassung geht hervor, bei welchen

Variablen der höchste Diskriminationswert vorliegt. Beispiel: Betrachten Sie die Tabelle der Mehrfachanpassung für *Marital status* und *Newspaper read most often*. Die Summe der Anpassungswerte über die beiden Dimensionen ist gleich 1,122 für *Marital status* und 0,911 für *Newspaper read most often*. Diese Information bedeutet, dass der Familienstand einer Person einen höheren Diskriminationswert liefert als die abonnierte Zeitung.

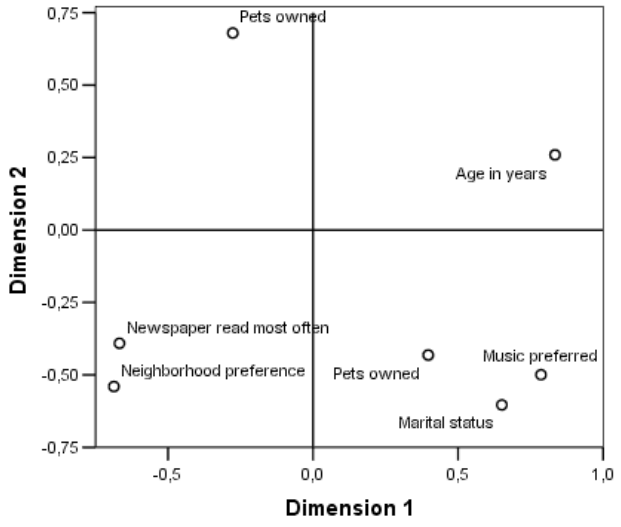
Die Einzelanpassung entspricht der quadrierten Gewichtung für die einzelnen Variablen und ist gleich der Varianz bei den Einzelkategoriekoordinaten. Die Gewichtungen sind daher gleich den Standardabweichungen der Einzelkategoriekoordinaten. Aus der Aufteilung der Einzelanpassung auf die verschiedenen Dimensionen geht hervor, dass die Diskrimination der Variablen *Newspaper read most often* hauptsächlich in der ersten Dimension erfolgt, die Diskrimination der Variablen *Marital status* dagegen nahezu vollständig in der zweiten Dimension. Die Kategorien von *Newspaper read most often* sind also in der ersten Dimension weiter voneinander entfernt als in der zweiten; für *Marital status* gilt dagegen der umgekehrte Fall. Im Gegensatz dazu findet die Diskrimination von *Age in years* sowohl in der ersten als auch in der zweiten Dimension statt; die Streuung der Kategorien ist daher in beiden Dimensionen gleich.

Komponentenladungen

Die nachstehende Abbildung zeigt das Diagramm der Komponentenladungen für die Untersuchungsdaten. Wenn keine Daten fehlen, sind die Komponentenladungen identisch mit den Pearson-Korrelationen zwischen den quantifizierten Variablen und den Objektwerten.

Die Distanz vom Ursprung zu den einzelnen Variablenpunkten ist etwa gleich der Wichtigkeit dieser Variable. Die kanonischen Variablen werden nicht in das Diagramm aufgenommen, können jedoch mit horizontalen und vertikalen Linien durch den Ursprung dargestellt werden.

Abbildung 11-17
Komponentenladungen



Die Beziehungen zwischen den Variablen sind offensichtlich. Zwei Richtungen fallen nicht mit der horizontalen oder vertikalen Achse zusammen. Eine Richtung wird durch *Age in years*, *Newspaper read most often* und *Neighborhood preference* bestimmt. Die andere Richtung ist

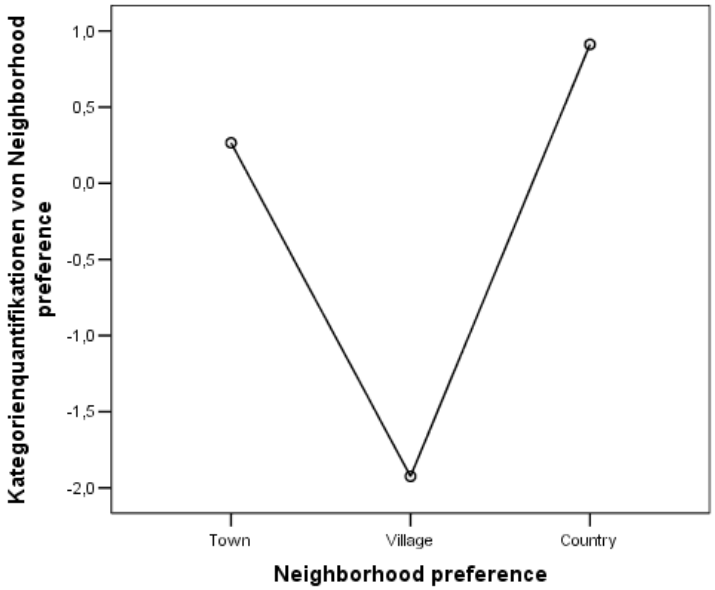
durch die Variablen *Marital status*, *Music preferred* und *Pets owned* definiert. Die Variable *Pets owned* ist eine mehrfach nominale Variable; hierfür werden also zwei Punkte aufgetragen. Die Quantifikationen werden jeweils als einzelne Variable interpretiert.

Transformationsdiagramme

Die verschiedenen Niveaus, auf denen die einzelnen Variablen skaliert werden können, bringen Einschränkungen für die Quantifizierungen mit sich. Transformationsdiagramme illustrieren die Beziehung zwischen den Quantifizierungen und den ursprünglichen Kategorien, die sich aus dem jeweils ausgewählten Messniveau der optimalen Skalierung ergibt.

Das Transformationsdiagramm für die Variable *Neighborhood preference*, die als nominal behandelt wurde, zeigt ein U-förmiges Muster, bei dem die mittlere Kategorie die niedrigste Quantifikation erhält, die äußeren Kategorien dagegen ähnliche Werte. Dieses Muster weist auf eine quadratische Beziehung zwischen der ursprünglichen Variablen und der transformierten Variablen hin. Ein alternatives Messniveau mit optimaler Skalierung für *Neighborhood preference* wird nicht empfohlen.

Abbildung 11-18
Transformationsdiagramm für die Variable Neighborhood preference (nominal)



Im Gegensatz dazu deuten die Quantifikationen für *Newspaper read most often* auf einen ansteigenden Trend in den drei Kategorien hin, in denen beobachtete Fälle vorliegen. Die erste Kategorie erhält die niedrigste Quantifikation, die zweite Kategorie den nächsthöheren Wert, und

die dritte Kategorie schließlich die höchste Quantifikation. Die Variable wird zwar als nominal skaliert; die Reihenfolge der Kategorien wird jedoch in den Quantifikationen beibehalten.

Abbildung 11-19
Transformationsdiagramm für Newspaper read most often (nominal)

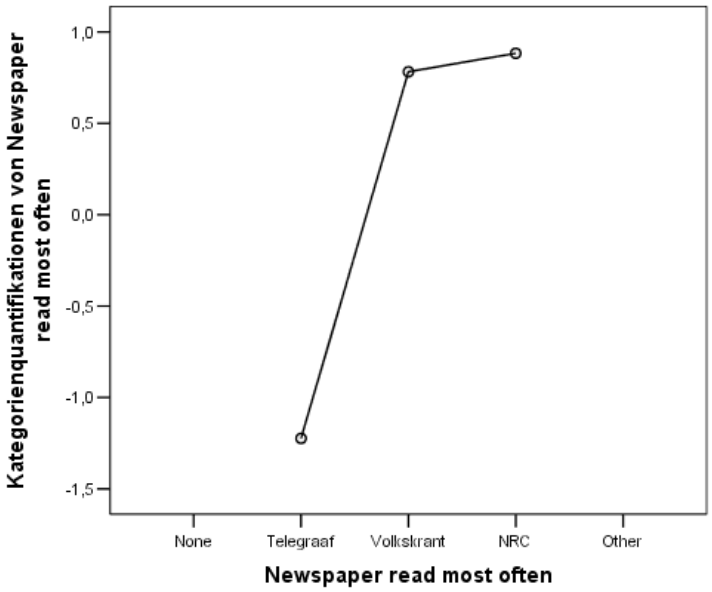
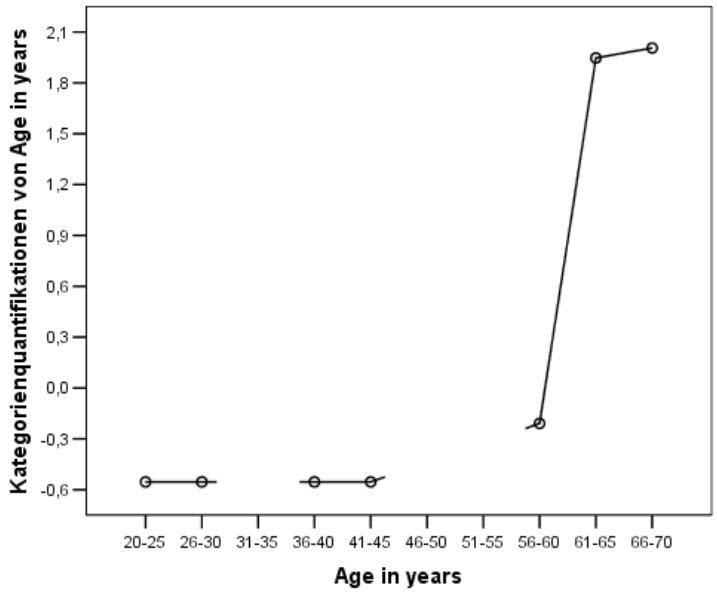


Abbildung 11-20
Transformationsdiagramm für "Age in years" (ordinal)



Das Transformationsdiagramm für *Age in years* zeigt eine S-förmige Kurve. Die vier beobachteten Kategorien mit dem niedrigsten Alter erhalten jeweils dieselbe negative Quantifikation, die beiden Kategorien mit dem höchsten Alter dagegen ähnliche positive Werte. In diesem Fall könnten die niedrigeren Altersgruppen in einer gemeinsamen Kategorie (z. B. unter 50) und die

beiden höheren Altersgruppen in einer zweiten Kategorie zusammengefasst werden. Die genaue Übereinstimmung der Quantifikationen für die niedrigeren Altersgruppen weist jedoch darauf hin, dass eine Einschränkung der Quantifikationen auf die Reihenfolge der ursprünglichen Kategorien ggf. nicht empfehlenswert wäre. Dei Quantifikationen für die Gruppen 26–30, 36–40 und 41–45 können nicht niedriger sein als die Quantifikation für die Gruppe 20–25. Diese Werte werden daher auf den Grenzwert gesetzt. Wenn Sie angeben, dass diese Werte kleiner sein können als die Quantifikation für die niedrigste Altersgruppe (also das Alter als nominal behandeln), kann dies die Anpassungsgüte erhöhen. Das Alter kann somit zwar als ordinale Variable angesehen werden; eine Behandlung als ordinale Variable ist in diesem Fall wohl eher nicht angebracht. Würde das Alter als numerisch behandelt werden (und so die Distanzen zwischen den Kategorien aufrechterhalten), würde dies zu einer beträchtlichen Verringerung der Anpassungsgüte führen.

Einzelkategoriekoordinaten im Vergleich mit Mehrfachkategoriekoordinaten

Bei jeder Variablen mit Behandlung als einfach nominal, ordinal oder numerisch werden Quantifikationen, Einzelkategoriekoordinaten und Mehrfachkategoriekoordinaten ermittelt. Diese Statistiken werden für *Age in years* dargestellt.

Abbildung 11-21
Koordinaten für "Age in years"

			Koordinaten von Einzelkategorien		Mehrere Kategorienkoordinaten	
			Dimension		Dimension	
	Randhäufigkeit	Quantifikation	1	2	1	2
20-25	3	-,554	-,377	-,437	-,192	-,139
26-30	5	-,554	-,377	-,437	-,404	-,623
31-35	0	,000				
36-40	1	-,554	-,377	-,437	-,318	-,733
41-45	1	-,554	-,377	-,437	-,356	-,534
46-50	0	,000				
51-55	0	,000				
56-60	2	-,209	-,142	-,165	-,435	,087
61-65	1	1,947	1,324	1,536	1,710	1,204
66-70	2	2,006	1,364	1,583	1,215	1,711
Fehlend	0					

Jede Kategorie, für die keine Fälle erfasst wurden, erhält die Quantifikation 0. Für *Age in years* umfasst dies die Kategorien 31–35, 46–50 und 51–55. Diese Kategorien müssen nicht gemeinsam mit dem anderen Kategorien eingeordnet werden und wirken sich nicht auf die Berechnungen aus.

Bei mehrfach nominalen Variablen erhalten die einzelnen Kategorien jeweils eine andere Quantifikation in den verschiedenen Dimensionen. Bei allen anderen Transformationstypen besitzt jede Kategorie nur eine einzige Quantifikation, unabhängig von der Anzahl der Dimensionen in der Lösung. Jedes Set von Einzelkategoriekoordinaten stellt die Position der Kategorie auf einer Linie im Objektraum dar. Die Koordinaten einer gegebenen Kategorie entsprechen der Quantifikation multipliziert mit der Dimensionsgewichtungen für die Variable. Beispiel: In der Tabelle für *Age in years* entsprechen die Einzelkategoriekoordinaten für die Kategorie 56-60 (-0,142; -0.165) der Quantifikation (-0,209) multipliziert mit den Gewichtungen (0,680; 0,789).

Die Mehrfachkategoriekoordinaten für Variablen, die als einzeln nominal, ordinal oder numerisch behandelt werden, stellen die Koordinaten der Kategorien im Objektraum vor dem Anwenden von ordinalen oder linearen Einschränkungen dar. Diese Werte sind

Verlustminimierungen ohne Nebenbedingungen. Bei mehrfach nominalen Variablen repräsentieren diese Koordinaten die Quantifikationen der Kategorien.

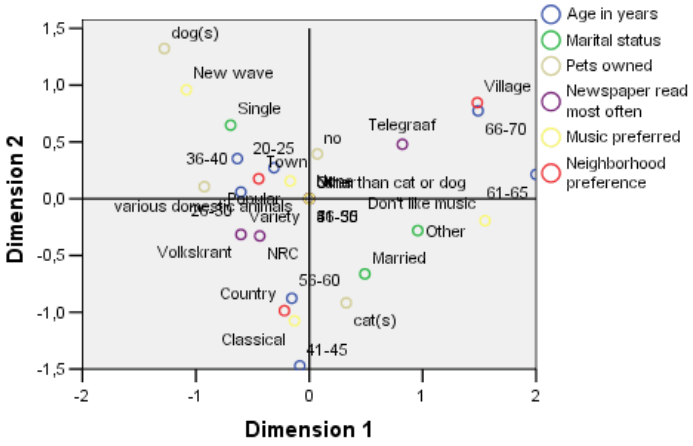
Der Vergleich der Einzelkategoriekoordinaten mit den Mehrfachkategoriekoordinaten zeigt, wie sich die Einschränkungen auf die Beziehung zwischen den Kategorien und deren Quantifikationen auswirken. In der ersten Dimension sinken die Mehrfachkategoriekoordinaten für *Age in years* in Kategorie 2 und bleiben bis Kategorie 9 etwa auf demselben Niveau; in dieser Kategorie erfolgt dann ein deutlicher Anstieg. Ein ähnliches Muster ist bei der zweiten Dimension zu sehen. Diese Beziehungen werden in den Einzelkategoriekoordinaten entfernt, bei denen die ordinale Einschränkung angewendet wird. In beiden Dimensionen findet nun kein Abfall der Koordinaten mehr statt. Die unterschiedliche Struktur der beiden Koordinaten-Sets legt nahe, dass ggf. eine nominale Behandlung angebrachter wäre.

Zentroide und projizierte Zentroide

Das Diagramm der Zentroide, die mit Variablen beschriftet sind, sollte auf dieselbe Weise interpretiert werden wie das Diagramm der Kategoriequantifikationen bei der Homogenitätsanalyse oder wie die Mehrfachkategoriekoordinaten bei der nichtlinearen Hauptkomponentenanalyse. Ein solches Diagramm zeigt klar, inwieweit die Variablen die Objektgruppen voneinander trennen. (Die Zentroiden befinden sich am Schwerpunkt der Objekte.)

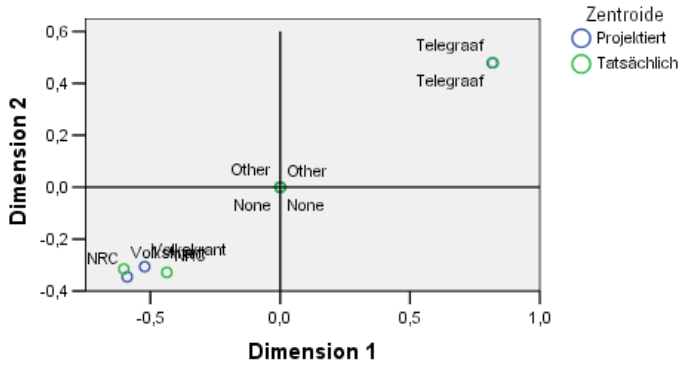
Beachten Sie, dass die Kategorien für *Age in years* nicht sehr deutlich getrennt sind. Die Kategorien mit den niedrigeren Altersgruppen sind im linken Bereich des Diagramms gruppiert. Wie bereits erwähnt wurde, ist das Messniveau mit ordinaler Skalierung für *Age in years* wahrscheinlich zu streng.

Abbildung 11-22
Zentroide, mit Variablen beschriftet



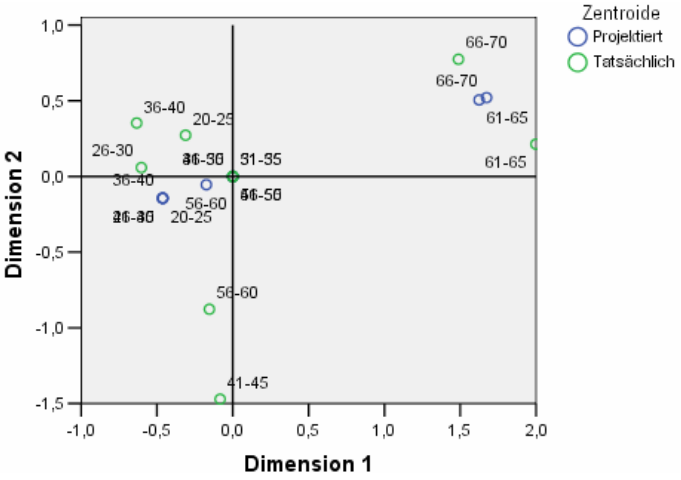
Wenn Sie Diagramme der Zentroide anfordern, werden auch Diagramme mit einzelnen Zentroiden und mit projizierten Zentroiden für jede Variable erstellt und jeweils mit den Wertelabels beschriftet. Die projizierten Zentroide befinden sich auf einer Linie im Objektraum.

Abbildung 11-23
Zentroide und projizierte Zentroide für "Newspaper read most often"



Die eigentlichen Zentroide werden auf die Vektoren projiziert, die durch die Komponentenladungen definiert sind. Diese Vektoren wurden in die Diagramme der Zentroide aufgenommen, um so leichter zwischen den projizierten Zentroiden und den eigentlichen Zentroiden unterscheiden zu können. Die projizierten Zentroide fallen in einen von vier Quadranten, die aus der Verlängerung zweier lotrechter Referenzlinien durch den Ursprung gebildet werden. Die Interpretation der Richtung von einzeln nominalen, ordinalen oder numerischen Variablen ergibt sich aus der Position der projizierten Zentroide. Beispiel: Die Variable *Newspaper read most often* ist als einzeln nominal angegeben. Die projizierten Zentroide zeigen, dass *Volkskrant* und *NRC* im Gegensatz zu *Telegraaf* stehen.

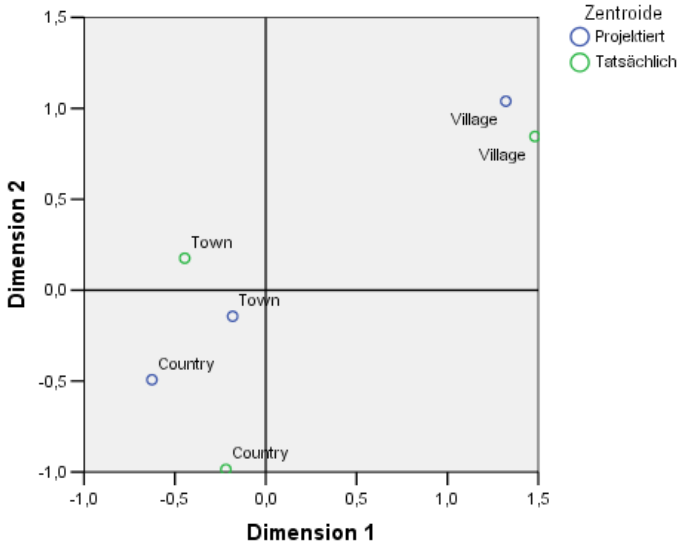
Abbildung 11-24
Zentroide und projizierte Zentroide für "Age in years"



Das Problem bei *Age in years* zeigt sich deutlich in den projizierten Zentroiden. Wird *Age in years* als ordinal behandelt, impliziert dies, dass die Reihenfolge der Altersgruppen beibehalten werden muss. Um diese Einschränkung zu erfüllen, werden alle Altersgruppen bis 45 Jahre auf denselben

Punkt projiziert. Entlang der Richtung, die durch *Age in years*, *Newspaper read most often* und *Neighborhood preference* definiert ist, findet keine Trennung der niedrigeren Altersgruppen statt. Diese Erkenntnis legt nahe, dass die Variable als nominal behandelt werden sollte.

Abbildung 11-25
Zentroide und projizierte Zentroide für "Neighborhood preference"



Um die Beziehungen zwischen den Variablen zu erkennen, ermitteln Sie, welche Kategorien (Werte) für Cluster von Kategorien in den Diagrammen mit Zentroiden stehen. Die Beziehungen zwischen *Age in years*, *Newspaper read most often* und *Neighborhood preference* können mit einem Blick in den oberen rechten und den unteren linken Bereich der Diagramme beschrieben werden. Die Altersgruppen im oberen rechten Bereich umfassen die älteren Teilnehmer; diese Personen lesen den Telegraaf und wohnen lieber in einem Dorf. Aus dem unteren linken Bereich der Diagramme geht hervor, dass die jungen Teilnehmer und die Teilnehmer mittleren Alters den Volkskrant oder NRC lesen und auf dem Land oder in einer Stadt wohnen möchten. Eine Trennung der niedrigeren Altersgruppen ist allerdings sehr schwierig.

Dieselben Interpretationen können für die andere Richtung (*Music preferred*, *Marital status* und *Pets owned*) vorgenommen werden; hierbei werden entsprechend der obere linke und der untere rechte Bereich des Diagramms mit den Zentroiden betrachtet. Aus dem oberen linken Bereich geht hervor, dass Alleinstehende häufig einen Hund halten und gern New Wave hören. Verheiratete und Angehörige anderer Kategorien beim Familienstand halten Katzen, wobei Verheiratete gern klassische Musik hören und die anderen Personen keine Musik mögen.

Eine alternative Analyse

Die Ergebnisse der Analyse legen nahe, dass die Behandlung der Variablen *Age in years* als ordinal nicht ganz angemessen ist. *Age in years* wird zwar auf ordinalem Messniveau gemessen; die Beziehungen mit anderen Variablen sind jedoch nicht monoton. Um die Auswirkungen zu untersuchen, die beim Wechsel zum einzeln nominalen Messniveau mit optimaler Skalierung entstehen, können Sie die Analyse wiederholen.

So führen Sie die Analyse durch:

- ▶ Rufen Sie das Dialogfeld "Nichtlineare kanonische Korrelationsanalyse" erneut auf und wechseln Sie zum ersten Set.
- ▶ Wählen Sie *age* aus und klicken Sie auf Bereich und Skala definieren.
- ▶ Wählen Sie im Dialogfeld "Bereich und Skala definieren" den Skalierungsbereich Einzeln nominal aus.
- ▶ Klicken Sie auf Weiter.
- ▶ Klicken Sie im Dialogfeld "Nichtlineare kanonische Korrelationsanalyse" auf OK.

Die Eigenwerte für eine zweidimensionale Lösung sind gleich 0,806 bzw. 0,757; die Gesamtanpassung beträgt 1,564.

Abbildung 11-26
Eigenwerte für die zweidimensionale Lösung

		Dimension		
		1	2	Summe
Verlustfunktion	Set 1	,249	,115	,363
	Set 2	,176	,408	,584
	Set 3	,157	,205	,363
	Mittelwert	,194	,243	,436
Eigenwert		,806	,757	
Anpassung				1,564

Die Tabellen für die Mehrfachanpassung und Einzelanpassung zeigen, dass *Age in years* immer noch einen hohen Diskriminationswert aufweist, wie die Summe der Werte für die Mehrfachanpassung belegt. Im Gegensatz zu den früheren Ergebnissen zeigt die Untersuchung der Einzelanpassungswerte jedoch, dass die Diskrimination nahezu vollständig in der zweiten Dimension erfolgt.

Abbildung 11-27
Aufteilen der Anpassung und der Verluste

		Mehrfachanpassung			Einzelne Anpassung			Einzelne Verlustfunktion		
		Dimension			Dimension			Dimension		
Set		1	2	Summe	1	2	Summe	1	2	Summe
1	Age in years[a]	,246	1,197	1,443	,195	1,188	1,384	,051	,008	,059
	Marital status[a]	,273	1,136	1,409	,272	1,135	1,407	,001	,000	,002
2	Pets owned[b]	,530	,392	,921						
	Newspaper read most often[a]	,639	,185	,824	,631	,149	,780	,008	,036	,044
3	Music preferred[a]	,604	,438	1,041	,603	,437	1,040	,000	,001	,001
	Neighborhood preference[a]	,075	,822	,897	,075	,822	,897	,000	,000	,000

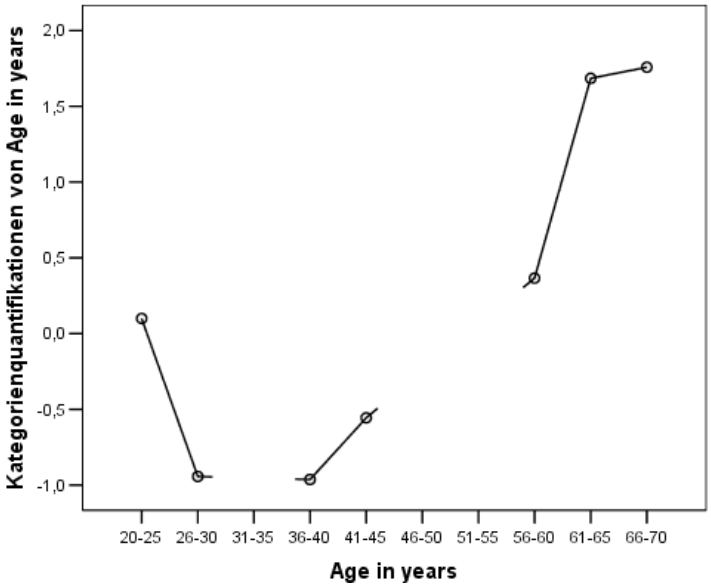
a. Optimales Skalierungsniveau: Einfach nominal

b. Optimales Skalierungsniveau: mehrfach nominal

Betrachten Sie das Transformationsdiagramm für *Age in years*. Die Quantifikationen für eine nominale Variable sind unbeschränkt; der nicht abnehmende Trend, der bei der ordinalen Behandlung von *Age in years* vorlag, ist daher nicht mehr vorhanden. Bis zum Alter von 40 Jahren besteht ein abnehmender Trend, ab diesem Alter liegt ein ansteigender Trend vor. Dies

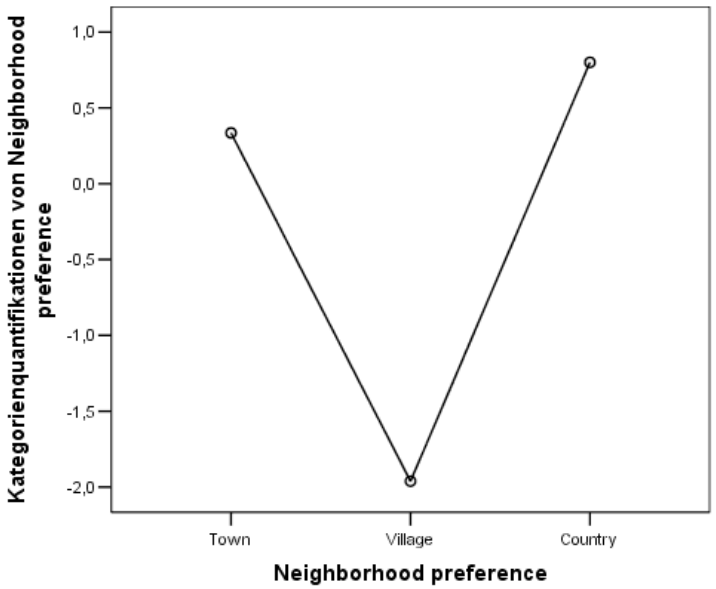
entspricht einer U-förmigen (quadratischen) Beziehung. Die beiden Kategorien mit den höheren Altersgruppen erhalten weiterhin ähnliche Werte und in nachfolgenden Analysen können diese Kategorien ggf. kombiniert werden.

Abbildung 11-28
Transformationsdiagramm für "Age in years" (nominal)



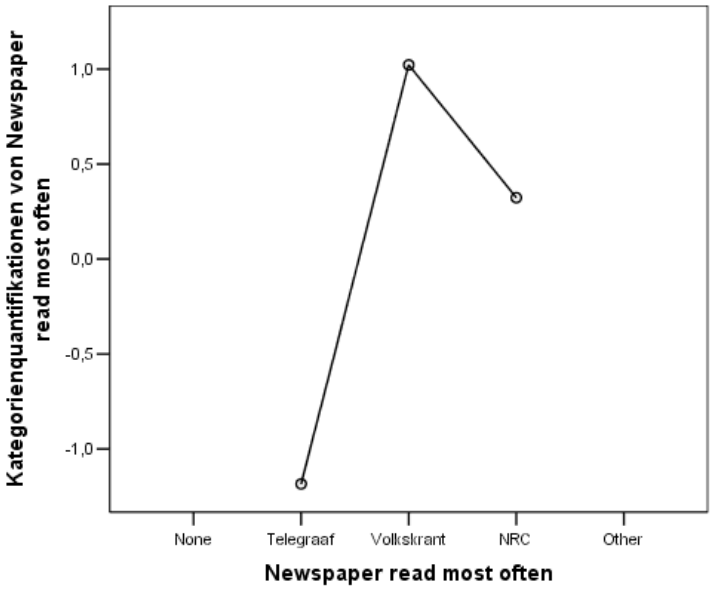
Hier wird das Transformationsdiagramm für *Neighborhood preference* dargestellt. Die nominale Behandlung von *Age in years* wirkt sich nicht nennenswert auf die Quantifikationen für *Neighborhood preference* aus. Die mittlere Kategorie erhält die niedrigste Quantifikation, die äußeren Kategorien dagegen große positive Werte.

Abbildung 11-29
Transformationsdiagramm für die Variable Neighborhood preference ("Alter" nominal)



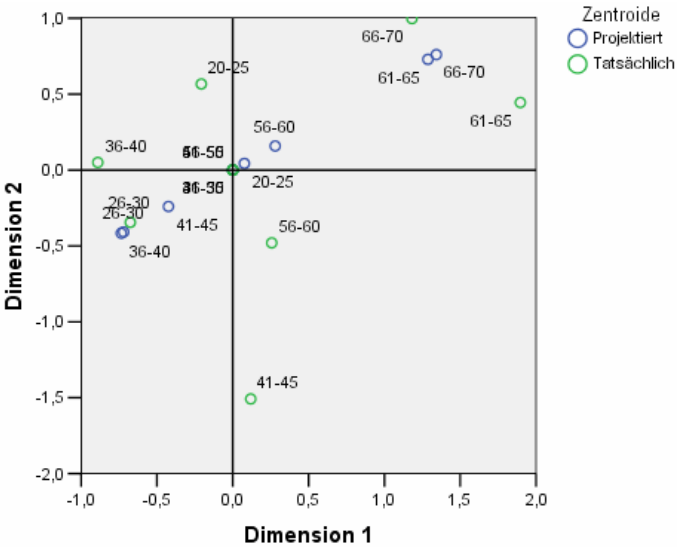
Beim Transformationsdiagramm für *Newspaper read most often* hat sich eine Änderung ergeben. Bislang bestand ein ansteigender Trend in den Quantifikationen, der unter Umständen eine ordinale Behandlung dieser Variablen nahelegt. Durch die nominale Behandlung von *Age in years* wurde dieser Trend jedoch aus den Quantifikationen für news entfernt.

Abbildung 11-30
Transformationsdiagramm für Newspaper read most often ("Alter", nominal)



Dies ist das Diagramm der Zentroide für *Age in years*. Beachten Sie, dass die Kategorien entlang der Linie, die die projizierten Zentroide verbindet, nicht in chronologischer Reihenfolge vorliegen. Die Gruppe "20–25" befindet sich nicht am Ende, sondern in der Mitte. Die Streuung der Kategorien ist im Vergleich zur bisherigen ordinalen Behandlung deutlich besser.

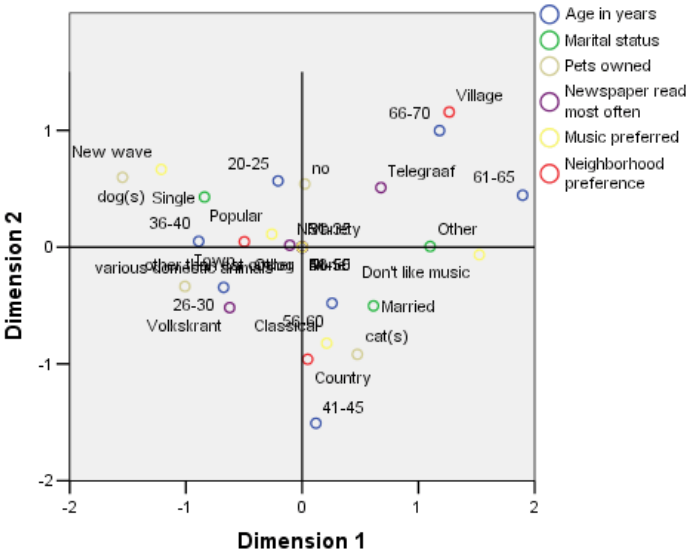
Abbildung 11-31
Zentroide und projizierte Zentroide für "Age in years" (nominal)



Das Diagramm mit den Zentroiden ermöglicht nun die Interpretation der niedrigeren Altersgruppen. Außerdem sind die Kategorien *Volkskrant* und *NRC* weiter voneinander entfernt als in der bisherigen Analyse; so wird eine separate Interpretation dieser Elemente möglich. Die Gruppen im Alter von 26 und 45 lesen den Volkskrant und leben gern auf dem Land. Die Altersgruppen 20–25 und 56–60 lesen NRC, wobei die jüngere Gruppe bevorzugt in der Stadt lebt und die ältere Gruppe lieber auf dem Land. Die Gruppen mit dem höchsten Alter lesen den Telegraaf und wohnen gern in einem Dorf.

Die Interpretation der anderen Richtung (*Music preferred*, *Marital status* und *Pets owned*) ist im Vergleich zur vorangegangenen Analyse praktisch unverändert. Der einzige offensichtliche Unterschied ist, dass Personen mit dem Familienstand ("Marital Status") *Other* entweder eine Katze halten oder gar kein Haustier besitzen.

Abbildung 11-32
Zentroide, mit Variablen beschriftet ("Alter" nominal)



Allgemeine Vorschläge

Nach der Untersuchung der ersten Ergebnisse soll die Analyse in der Regel weiter verfeinert werden, indem Sie einen Teil der Angaben für die nichtlineare kanonische Korrelationsanalyse ändern. Hier finden Sie einige Tips, wie Sie die Analyse strukturieren:

- Erstellen Sie so viele Sets wie möglich. Wichtige Variablen für die Vorhersage sollten dabei jeweils in ein separates Set eingebracht werden.
- Variablen, die Ihrer Ansicht nach als Einflußgrößen fungieren, sollten gemeinsam in ein Set plaziert werden. Falls zahlreiche Einflußgrößen vorliegen, teilen Sie sie nach Möglichkeit auf mehrere Sets auf.

- Mehrfach nominale Variablen sollten jeweils in ein separates Set aufgenommen werden.
- Wenn Variablen hoch korreliert sind und diese Beziehung die Lösung nicht dominieren soll, platzieren Sie die betreffenden Variablen in einem einzigen Set.

Empfohlene Literatur

In folgenden Texten finden Sie weitere Informationen zur nichtlinearen kanonischen Korrelationsanalyse:

Carroll, J. D. 1968. Generalization of canonical correlation analysis to three or more sets of variables. In: *Proceedings of the 76th Annual Convention of the American Psychological Association, 3,* Washington, D.C.: American Psychological Association, 227–228.

De Leeuw, J. 1984. *Canonical analysis of categorical data*, 2nd (Hg.). Leiden: DSWO Press.

Horst, P. 1961. Generalized canonical correlations and their applications to experimental data. *Journal of Clinical Psychology*, 17, 331–347.

Horst, P. 1961. Relations among m sets of measures. *Psychometrika*, 26, 129–149.

Kettenring, J. R. 1971. Canonical analysis of several sets of variables. *Biometrika*, 58, 433–460.

Van der Burg, E. 1988. *Nonlinear canonical correlation and some related techniques*. Leiden: DSWO Press.

Van der Burg, E., als auch J. De Leeuw. 1983. Nonlinear canonical correlation. *British Journal of Mathematical and Statistical Psychology*, 36, 54–80.

Van der Burg, E., J. De Leeuw, als auch R. Verdegaal. 1988. Homogeneity analysis with k sets of variables: An alternating least squares method with optimal scaling features. *Psychometrika*, 53, 177–197.

Verboon, P., als auch I. A. Van der Lans. 1994. Robust canonical discriminant analysis. *Psychometrika*, 59, 485–507.

Kapitel
12

Korrespondenzanalyse

Eine **Korrespondenztabelle** ist jede Zweifach-Tabelle, deren Zellen eine gewisse Korrespondenz zwischen den Zeilen und Spalten zeigen. Das Korrespondenzmaß kann ein Hinweis auf die Ähnlichkeit, Affinität, Konfusion, Assoziation oder Wechselwirkung zwischen den Zeilen- und Spaltenvariablen sein. Ein sehr häufiger Typ einer Korrespondenztabelle ist eine Kreuztabelle, bei der die Zellen Häufigkeitszählungen enthalten.

Diese Tabellen können schnell und einfach mit der Prozedur "Kreuztabellen" erstellt werden. Eine Kreuztabelle gibt den Charakter der Beziehung zwischen den beiden Variablen jedoch nicht immer klar wieder. Dies gilt insbesondere dann, wenn die relevanten Variablen nominal sind (ohne natürliche Reihenfolge oder natürlichen Rang) und zahlreiche Kategorien besitzen. Eine Kreuztabelle kann zwar aussagen, dass die beobachteten Zellhäufigkeiten deutlich von den erwarteten Werten in einer $10x9$-Kreuztabelle von *Berufsgruppe* und *Frühstücksflocken* abweichen. Es kann jedoch sehr schwierig zu bestimmen sein, welche Berufsgruppen einen ähnlichen Geschmack aufweisen oder welcher genau dies ist.

Mit einer Korrespondenzanalyse können Sie die Beziehung zwischen zwei nominalen Variablen grafisch in einem mehrdimensionalen Raum untersuchen. Hierbei werden Zeilen- und Spaltenwerte berechnet und Diagramme auf der Grundlage dieser Werte angelegt. Ähnliche Kategorien werden in den Diagrammen nahe beieinander dargestellt. Auf diese Weise ist leicht ersichtlich, welche Kategorien einer Variablen einander ähneln oder welche Kategorien beider Variablen zusammenhängen. Bei der Korrespondenzanalyse ist es außerdem möglich, Zusatzpunkte in den Raum einzugliedern, der durch die aktiven Punkte definiert ist.

Falls die Reihenfolge der Kategorien gemäß den Werten nicht erwünscht oder nicht intuitiv ist, stehen Ordnungs-Einschränkungen zur Verfügung, mit denen Sie festlegen können, dass die Werte für einige Kategorien gleich sein müssen. Beispiel: Bei der Variablen *Rauchgewohnheit* mit den Kategorien *Nichtraucher*, *Leicht*, *Mittel* und *Stark* werden Werte erwartet, die dieser Reihenfolge entsprechen. Werden die Kategorien in der Analyse jedoch als *Nichtraucher*, *Leicht*, *Stark* und *Mittel* geordnet, können Sie die Werte für *Stark* und *Mittel* gleichsetzen und so die Reihenfolge der Kategorien in den Werten beibehalten.

Die Interpretation der Korrespondenzanalyse hinsichtlich der Distanzen ist abhängig von der verwendeten Normalisierungsmethode. Mit der Korrespondenzanalyse können Sie wahlweise die Differenzen zwischen den Kategorien einer Variablen oder die Differenzen zwischen mehreren Variablen untersuchen. Bei der Standard-Normalisierung werden die Differenzen zwischen den Zeilen- und Spaltenvariablen analysiert.

Der Algorithmus für die Korrespondenzanalyse ermöglicht viele verschiedene Spielarten. Bei der Standard-Korrespondenzanalsye werden die Zeilen und Spalten zentriert, und es werden Chi-Quadrat-Distanzen herangezogen. Wenn Sie jedoch alternative Zentrierungsoptionen in Kombination mit Euklidischen Distanzen verwenden, entsteht eine alternative Repräsentation einer Matrix in einem flachdimensionierten Raum.

Es werden drei Beispiele vorgestellt. Das erste Beispiel umfaßt eine relativ kleine Korrespondenztabelle und zeigt die grundlegenden Konzepte dieser Analyse. Das zweite Beispiel zeigt die praktische Anwendung im Marketing-Bereich. Im letzten Beispiel wird eine Distanztabelle in einer mehrdimensionalen Skalierung verwendet.

Normalisierung

Bei der Normalisierung wird die Trägheit über die Zeilen- und Spaltenwerte gestreut. Einige Aspekte der Korrespondenzanalysenlösung (z. B. die Singulärwerte, die Trägheit pro Dimension und die Beiträge) bleiben bei den verschiedenen Normalisierungsmethoden unverändert. Die Zeilen- und Spaltenwerte und die zugehörigen Varianzen werden von der Änderung erfaßt. Die Trägheit wird bei der Korrespondenzanalyse auf verschiedene Weise gestreut. Die drei häufigsten Verfahren streuen die Trägheit nur über die Zeilenwerte, nur über die Spaltenwerte oder symmetrisch über Zeilen- und Spaltenwerte.

Zeilenprinzipal. Bei der Zeilenprinzipal-Normalisierung nähern sich die Euklidischen Distanzen zwischen den Zeilenpunkten den Chi-Quadrat-Distanzen zwischen den Zeilen der Korrespondenztabelle. Die Zeilenwerte entsprechen dem gewichteten Durchschnitt der Spaltenwerte. Die Spaltenwerte werden so standardisiert, dass eine gewichtete Summe der quadrierten Distanzen zum Zentroid 1 entsteht. Bei dieser Methode werden die Distanzen zwischen den Zeilenkategorien maximiert. Wenn Sie hauptsächlich ermitteln möchten, wie sich die Kategorien der Zeilenvariable voneinander unterscheiden, sollten Sie daher die Zeilenprinzipal-Normalisierung verwenden.

Spaltenprinzipal. Unter Umständen sollen die Chi-Quadrat-Distanzen zwischen den Spalten der Korrespondenztabelle genähert werden. In diesem Fall sollten die Spaltenwerte dem gewichteten Durchschnitt der Zeilenwerte entsprechen. Die Zeilenwerte werden so standardisiert, dass eine gewichtete Summe der quadrierten Distanzen zum Zentroid 1 entsteht. Bei dieser Methode werden die Distanzen zwischen den Spaltenkategorien maximiert. Verwenden Sie diese Methode, wenn Sie hauptsächlich ermitteln möchten, inwieweit sich die Kategorien der Spaltenvariable voneinander unterscheiden.

Symmetrisch. Die Zeilen und Spalten können auch symmetrisch behandelt werden. Bei dieser Normalisierung wird die Trägheit gleichmäßig über die Zeilen- und Spaltenwerte gestreut. Beachten Sie in diesem Fall, dass weder die Distanzen zwischen den Zeilenpunkten noch die Distanzen zwischen den Spaltenpunkten eine Näherung der Chi-Quadrat-Distanzen bilden. Verwenden Sie diese Methode, wenn Sie hauptsächlich die Unterschiede oder Ähnlichkeiten zwischen den beiden Variablen untersuchen möchten. Diese Methode wird in der Regel für Biplots herangezogen.

Prinzipal. Bei einer vierten Option, der Prinzipal-Normalisierung, wird die Trägheit zweimal in der Lösung gestreut: zunächst über die Zeilenwerte und dann über die Spaltenwerte. Verwenden Sie diese Methode, wenn Sie die Distanzen zwischen den Zeilenpunkten und die Distanzen zwischen den Spaltenpunkten ermitteln möchten, sofern der Zusammenhang zwischen den Zeilen- und Spaltenpunkten dabei nicht relevant ist. Biplots sind für diese Normalisierungsoption nicht geeignet und stehen daher auch nicht zur Verfügung, wenn Sie die Prinzipal-Normalisierungsmethode auswählen.

Beispiel: Rauchgewohnheit nach Berufskategorie

Das Ziel der Korrespondenzanalyse ist es, die Beziehungen zwischen den Zeilen und Spalten einer Korrespondenztabelle aufzuzeigen. Die grundlegenden Konzepte werden anhand einer von Greenacre (Greenacre , 1984) vorgestellten hypothetischen Tabelle erläutert. Diese Informationen finden Sie in der Datei *smoking.sav*. Für weitere Informationen siehe Beispieldateien in Anhang A auf S. 326.

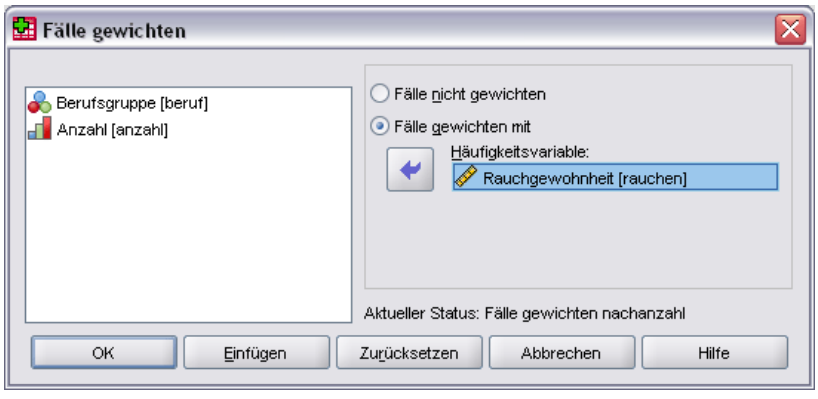
Die relevante Tabelle wird durch eine Kreuztabelle der Rauchgewohnheiten und der Berufskategorie gebildet. Die Variable *Berufsgruppe* enthält die Berufskategorien *Senior Manager*, *Junior Manager*, *Angestellter mit Erfahrung*, *Angestellter ohne Erfahrung* und *Sekretariat*, die bei der Bildung der Lösung herangezogen werden, außerdem die Kategorie *National Average*, die als Ergänzung der Analyse dienen kann. Die Variable *Rauchen* enthält die Rauchgewohnheiten *Nichtraucher*, *Leicht*, *Mittel* und *Stark*, die bei der Bildung der Lösung herangezogen werden, außerdem die Kategorien *No Alcohol* und *Alcohol*, die als Ergänzung der Analyse dienen können.

Durchführung der Analyse

- Bevor Sie die Prozedur "Korrespondenzanalyse" vornehmen können, müssen die Fälle gemäß dem Aufbau der Daten nach der Variablenanzahl gewichtet werden. Wählen Sie hierzu die folgenden Befehle aus den Menüs aus:

  Daten
  Fälle gewichten...

Abbildung 12-1
Dialogfeld "Fälle gewichten"



- Fälle gewichten nach *Anzahl*.
- Klicken Sie auf OK.

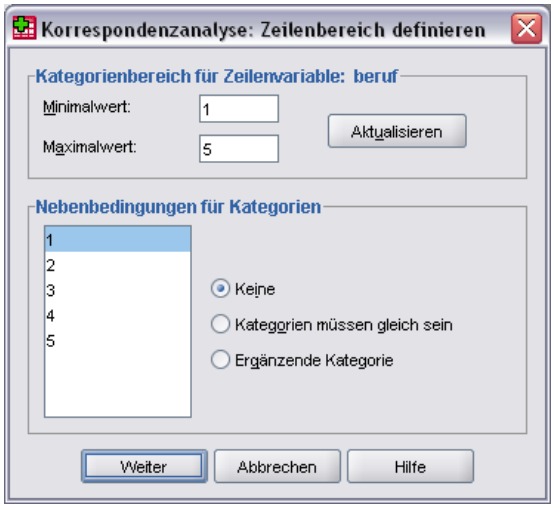
- Zum Erstellen einer Korrespondenzanalyse in zwei Dimensionen mit der Zeilenprinzipal-Normalisierung wählen Sie dann die folgenden Befehle aus den Menüs aus:

  Analysieren
  Dimensionsreduktion
  Korrespondenzanalyse...

Abbildung 12-2
Dialogfeld "Korrespondenzanalyse"

- Wählen Sie *Berufsgruppe* als Zeilenvariable aus.
- Klicken Sie auf Bereich definieren.

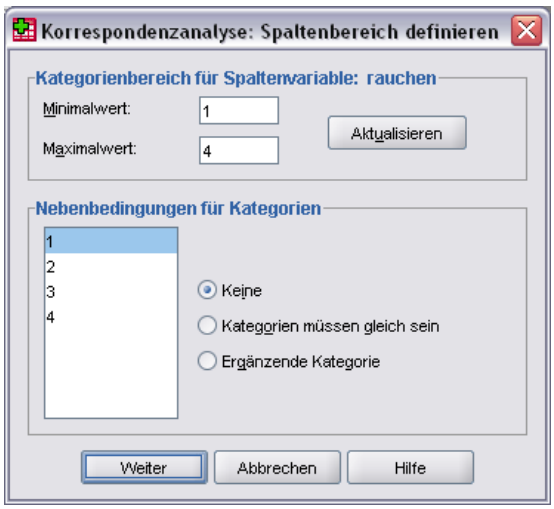
Abbildung 12-3
Dialogfeld "Zeilenbereich definieren"



- Geben Sie 1 als Mindestwert ein.
- Geben Sie 5 als Höchstwert ein.
- Klicken Sie auf Aktualisieren.
- Klicken Sie auf Weiter.
- Wählen Sie *Rauchen* als Spaltenvariable aus.

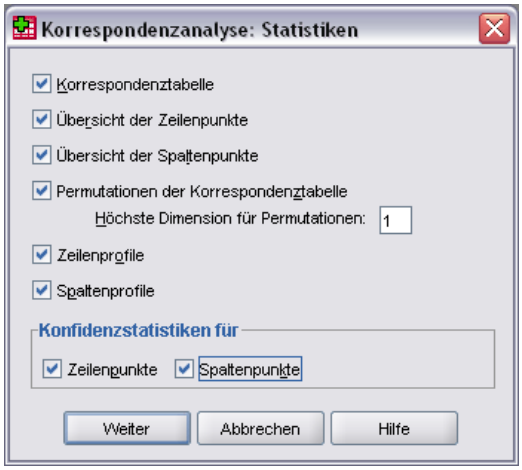
- Klicken Sie im Dialogfeld "Korrespondenzanalyse" auf Bereich definieren.

Abbildung 12-4
Dialogfeld "Spaltenbereich definieren"



- Geben Sie 1 als Mindestwert ein.
- Geben Sie 4 als Höchstwert ein.
- Klicken Sie auf Aktualisieren.
- Klicken Sie auf Weiter.
- Klicken Sie im Dialogfeld "Korrespondenzanalyse" auf Statistiken.

Abbildung 12-5
Dialogfeld "Statistiken"



- Wählen Sie Zeilenprofile und Spaltenprofile.
- Wählen Sie Permutationen der Korrespondenztabelle.
- Wählen Sie Konfidenzstatistiken für Zeilenpunkte und Spaltenpunkte.

- Klicken Sie auf Weiter.
- Klicken Sie im Dialogfeld "Korrespondenzanalyse" auf OK.

Korrespondenztabelle.

Die Korrespondenztabelle zeigt die Verteilung der Rauchgewohnheiten für fünf Ebenen der Berufskategorie. Die Zeilen der Korrespondenztabelle stehen für die Berufskategorien. Die Spalten entsprechen den Rauchgewohnheiten.

Abbildung 12-6
Korrespondenztabelle

	Rauchgewohnheit				
Berufsgruppe	Nichtraucher	Leicht	Mittel	Stark	Aktiver Rand
Senior Manager	4	2	3	2	11
Junior Manager	4	3	7	4	18
Angestellter mit Erfahrung	25	10	12	4	51
Angestellter ohne Erfahrung	18	24	33	13	88
Sekretariat	10	6	7	2	25
Aktiver Rand	61	45	62	25	193

Die Zeilenrandsummen zeigen, dass das Unternehmen deutlich mehr Angestellte (mit und ohne Erfahrung) als Manager und Sekretariatsangehörige beschäftigt. Die Aufteilung der Manager in Senior und Junior Manager entspricht jedoch in etwa der Aufteilung der Angestellten in Personen mit und ohne Erfahrung. Aus den Spaltensummen geht hervor, dass Nichtraucher und mittelschwere Raucher in etwa gleich häufig anzutreffen sind. Des weiteren sind die starken Raucher gegenüber den anderen drei Kategorien deutlich in der Minderheit. Haben diese Berufskategorien jedoch etwas bei den Rauchgewohnheiten gemeinsam und wenn ja, was? Und welche Beziehung besteht zwischen Berufsgruppe und Rauchen?

Dimensionen

Im Idealfall sollte eine Korrespondenzanalysenlösung erstellt werden, die die Beziehung zwischen den Zeilen- und Spaltenvariablen in so wenigen Dimensionen wie möglich zeigt. Häufig empfiehlt es sich jedoch, die Höchstanzahl der Dimensionen zu betrachten, um so den relativen Beitrag der einzelnen Dimensionen zu erkennen. Die maximal zulässige Anzahl an Dimensionen für eine Korrespondenzanalysenlösung entspricht der Anzahl der aktiven Zeilen minus 1 bzw. der Anzahl der aktiven Spalten minus 1 (niedrigerer Wert maßgeblich). Eine Zeile oder Spalte gilt als aktiv, wenn eine eindeutige Wertegruppe für diese Zeile bzw. Spalte existiert. Ergänzende Zeilen oder Spalten sind nicht aktiv. Im vorliegenden Beispiel ist die maximale Anzahl an Dimensionen gleich min(5,4) – 1 = 3.

Die erste Dimension zeigt den größtmöglichen Anteil an Trägheit (Messwert für die Variation in den Daten), die zweite Dimension verläuft orthogonal zur ersten und zeigt den größtmöglichen Anteil an der verbleibenden Trägheit usw. Die Gesamtträgheit kann in Komponenten aufgeteilt und dann den einzelnen Dimensionen zugeordnet werden. Anschließend können Sie die Trägheit

einer bestimmten Dimension mit der Gesamtträgheit vergleichen und so die Dimensionsträgheit auswerten. Beispiel: Die erste Dimension zeigt 87,8 % (0,075/0,085) der Gesamtträgheit, die zweite Dimension dagegen nur 11,8 % (0,010/0,085).

Abbildung 12-7
Trägheit pro Dimension

Dimension	Singulärwert	Auswertung für Trägheit	Chi-Quadrat	Sig.	Anteil der Trägheit		Singulärwert für Konfidenz	
					Bedingen	Kumuliert	Standardabweichung	Korrelation 2
1	,273	,075			,878	,878	,070	,020
2	,100	,010			,118	,995	,076	
3	,020	,000			,005	1,000		
Gesamt auswertung		,085	16,442	,172[a]	1,000	1,000		

a. 12 Freiheitsgrade

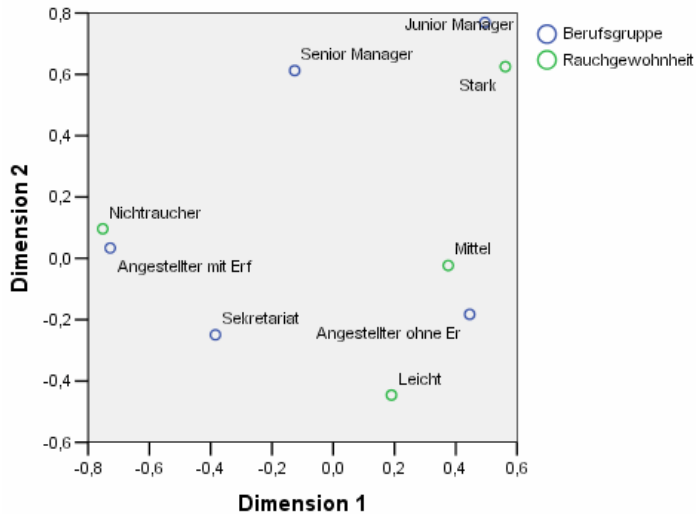
Wenn Sie feststellen, dass die ersten *p* Dimensionen einer *q*-dimensionalen Lösung einen ausreichenden Teil der Trägheit zeigen, können Sie höhere Dimensionen außer Acht lassen. In diesem Beispiel reicht die zweidimensionale Lösung völlig aus, weil die dritte Dimension weniger als 1,0 % der Gesamtträgheit repräsentiert.

Die Singulärwerte können als die Korrelation zwischen den Zeilen- und Spaltenwerten interpretiert werden. Diese Werte entsprechen dem Korrelationskoeffizienten nach Pearson (*r*) in der Korrelationsanalyse. In jeder Dimension ist der quadrierte Singulärwert (Eigenwert) gleich der Trägheit und bildet so einen weiteren Messwert für die Wichtigkeit der betreffenden Dimension.

Biplot

Bei der Korrespondenzanalyse werden verschiedene Diagramme erzeugt, die die zugrunde liegenden Beziehungen zwischen den Kategorien und zwischen den Variablen darstellen. Dies ist das Streudiagramm der Zeilen- und Spaltenwerte für die zweidimensionale Lösung.

Abbildung 12-8
Diagramm der Zeilen- und Spaltenwerte (symmetrische Normalisierung)



Die Interpretation des Diagramms ist unkompliziert: Zeilen-/Spaltenpunkte, die nahe beieinander liegen, sind einander ähnlicher als Punkte, die weit voneinander entfernt sind. Mit der zweiten Dimension werden Manager von anderen Angestellten getrennt, mit der ersten Dimension dagegen Senior-Kräfte von Junior-Kräften. Die Sekretariatsangehörigen befinden sich zwischen diesen Positionen.

Durch die symmetrische Normalisierung kann die Beziehung zwischen Berufsgruppe und Rauchgewohnheiten mühelos untersucht werden. Manager liegen beispielsweise näher an der Rauch-Kategorie *Stark*, Angestellte mit Erfahrung dagegen nahe an der Kategorie *Nichtraucher*. Angestellte ohne Erfahrung fallen offensichtlich in die Kategorien *Mittel* oder *Leicht*. Den Sekretariatsangehörigen lassen sich keine bestimmten Rauchgewohnheiten zuordnen; diese Mitarbeiter sind jedoch weit von der Kategorie *Stark* entfernt.

Profile und Distanzen

Um die Distanz zwischen den Kategorien zu ermitteln, werden bei der Korrespondenzanalyse nicht nur die Randverteilungen berücksichtigt, sondern auch die einzelnen Zellhäufigkeiten. Es werden Zeilen- und Spaltenprofile berechnet, also die Zeilen- und Spaltenanteile für jede Zelle auf der Grundlage der Randsummen.

Abbildung 12-9
Zeilenprofile (symmetrische Normalisierung)

	Rauchgewohnheit				
Berufsgruppe	Nichtraucher	Leicht	Mittel	Stark	Aktiver Rand
Senior Manager	,364	,182	,273	,182	1,000
Junior Manager	,222	,167	,389	,222	1,000
Angestellter mit Erfahrung	,490	,196	,235	,078	1,000
Angestellter ohne Erfahrung	,205	,273	,375	,148	1,000
Sekretariat	,400	,240	,280	,080	1,000
Masse	,316	,233	,321	,130	

Die Zeilenprofile geben den Anteil der Zeilenkategorie in den einzelnen Spaltenkategorien an. Die Angestellten mit Erfahrung sind beispielsweise zum größten Teil Nichtraucher; nur wenige Personen in dieser Gruppe sind starke Raucher. Unter den Junior Managern finden sich dagegen viele mittelschwere Raucher, jedoch nur sehr wenige leichte Raucher.

Die Spaltenprofile geben den Anteil der Spalte in den einzelnen Zeilenkategorien an. Die meisten leichten Raucher sind beispielsweise Angestellte ohne Erfahrung. Auch die meisten mittleren und starken Raucher fallen in die Gruppe der Angestellten ohne Erfahrung. Denken Sie daran, dass die Stichprobe zum größten Teil Angestellte ohne Erfahrung enthält. Es ist also nicht weiter überraschend, dass diese Berufskategorie in den Rauchkategorien vorherrscht.

Abbildung 12-10
Spaltenprofile

	Rauchgewohnheit				
Berufsgruppe	Nichtraucher	Leicht	Mittel	Stark	Masse
Senior Manager	,066	,044	,048	,080	,057
Junior Manager	,066	,067	,113	,160	,093
Angestellter mit Erfahrung	,410	,222	,194	,160	,264
Angestellter ohne Erfahrung	,295	,533	,532	,520	,456
Sekretariat	,164	,133	,113	,080	,130
Aktiver Rand	1,000	1,000	1,000	1,000	

Die Masse ist ein Messwert, der die Wichtigkeit eines Objekts auf der Grundlage der marginalen Häufigkeit angibt. Die Masse wirkt sich auf den **Zentroid** aus, also auf den gewichteten Mittelwert für das Zeilen- oder Spaltenprofil. Der Zeilenzentroid ist der Mittelwert für das Zeilenprofil. Punkte mit einer großen Masse (z. B. Angestellte ohne Erfahrung) ziehen den Zentroid stärker an ihre Position. Bei Punkten mit einer kleinen Masse (z. B. Senior Manager) wird der Zeilen-Zentroid entsprechend nur geringfügig zur Position gezogen.

Anders gesagt: Je größer die Differenz zwischen den Zeilenprofilen ist, desto größer ist die Distanz zwischen den Punkten im Diagramm. Wenn Sie beispielsweise die Zeilenprinzipal-Normalisierung verwenden, sind bei der endgültigen Konfiguration die Euklidischen Distanzen zwischen den Zeilenpunkten im volldimensionierten Raum gleich den Chi-Quadrat-Distanzen zwischen den Zeilen in der Korrespondenztabelle. In einem reduzierten Raum nähern sich die Euklidischen Distanzen den Chi-Quadrat-Distanzen. Die Chi-Quadrat-Distanzen bestehen wiederum aus gewichteten Profildistanzen. Diese gewichteten Distanzen beruhen auf der Masse.

Bei der Spaltenprinzipal-Normalisierung sind die Euklidischen Distanzen zwischen den Spaltenpunkten im volldimensionierten Raum gleich den Chi-Quadrat-Distanzen zwischen den Spalten der Korrespondenztabelle. Beachten Sie jedoch, dass diese Werte bei symmetrischer Normalisierung nicht gleich sind.

Die Gesamtträgheit ist definiert als die gewichtete Summe aller quadrierten Distanzen zum Ursprung dividiert durch die Summe über alle Zellen, wobei die Gewichtung anhand der Massen erfolgt. Zeilen mit geringer Masse beeinflussen die Trägheit nur dann, wenn sie weit entfernt vom Zentroiden sind. Zeilen mit großer Masse wirken sich auch dann auf die Gesamtträgheit aus, wenn sie sich dicht am Zentroid befinden. Dies gilt entsprechend für die Spalten.

Zeilen- und Spaltenwerte

Die Zeilen- und Spaltenwerte sind die Koordinaten der Zeilen- und Spaltenpunkte im Biplot.

Abbildung 12-11
Zeilenwerte (symmetrische Normalisierung)

		Wert in Dimension			Beitrag				
					des Punktes an der Trägheit der Dimension		der Dimension an der Trägheit des Punktes		
Berufsgruppe	Masse	1	2	Übersicht über Trägheit	1	2	1	2	Gesamt übersicht
Senior Manager	,057	-,126	,612	,003	,003	,214	,092	,800	,893
Junior Manager	,093	,495	,769	,012	,084	,551	,526	,465	,991
Angestellter mit Erfahrung	,264	-,728	,034	,038	,512	,003	,999	,001	1,000
Angestellter ohne Erfahrung	,456	,446	-,183	,026	,331	,152	,942	,058	1,000
Sekretariat	,130	-,385	-,249	,006	,070	,081	,865	,133	,999
Aktiver Gesamtwert	1,000			,085	1,000	1,000			

Abbildung 12-12
Spaltenwerte (symmetrische Normalisierung)

		Wert in Dimension			Beitrag				
					des Punktes an der Trägheit der Dimension		der Dimension an der Trägheit des Punktes		
Rauchgewohnheit	Masse	1	2	Übersicht über Trägheit	1	2	1	2	Gesamtübersicht
Nichtraucher	,316	-,752	,096	,049	,654	,029	,994	,006	1,000
Leicht	,233	,190	-,446	,007	,031	,463	,327	,657	,984
Mittel	,321	,375	-,023	,013	,166	,002	,982	,001	,983
Stark	,130	,562	,625	,016	,150	,506	,684	,310	,995
Aktiver Gesamtwert	1,000			,085	1,000	1,000			

a. Symmetrische Normalisierung

Die Spaltenwerte hängen über die Profile und den Singulärwert (aus der Tabelle mit der Trägheit pro Dimension) mit den Zeilenwerten zusammen. Genauer gesagt, die Zeilenwerte sind das Matrixprodukt der Zeilenprofile und der Spaltenwerte, skaliert mit dem Singulärwert

für jede Dimension. Der Wert –0,126 für Senior Manager in der ersten Dimension entspricht beispielsweise:

$$\frac{(0.364 \times -0.752) + (0.182 \times 0.190) + (0.273 \times 0.375) + (0.182 \times 0.562)}{0.273}$$

Bei der Zeilenprinzipal-Normalisierung wird der Singulärwert nicht in diese Gleichung aufgenommen. Die Zeilenpunkte befinden sich im gewichteten Zentroid der aktiven Spaltenpunkte, wobei die Gewichtungen den Einträgen in der Tabelle mit den Zeilenprofilen entsprechen. Wenn die Zeilenpunkte dem gewichteten Durchschnitt der Spaltenpunkte entsprechen und die maximal zulässige Anzahl an Dimensionen verwendet wird, ist die Euklidische Distanz zwischen einem Zeilenpunkt und dem Ursprung gleich der Chi-Quadrat-Distanz zwischen der Zeile und der Durchschnittszeile, die wiederum gleich der Trägheit einer Zeile ist. Die Chi-Quadrat-Statistik entspricht der Gesamtträgheit multipliziert mit der Summe aller Zellen in der Korrespondenztabelle. Die Ausrichtung der Zeilenpunkte kann daher als bildliche Darstellung der Chi-Quadrat-Statistik angesehen werden. Für die Spaltenprinzipal-Normalisierung gilt eine ähnliche Interpretation, nicht jedoch für die symmetrische Normalisierung.

Beiträge

Es ist möglich, die durch eine bestimmte Dimension gezeigte Trägheit zu berechnen. Die Werte in den einzelnen Dimensionen entsprechen einer orthogonalen Projektion von dem Punkt auf diese Dimension. Die Trägheit einer Dimension ist daher gleich der gewichteten Summe der quadrierten Distanzen von den Werten in der Dimension zum Ursprung. Ob diese Feststellung jedoch für Zeilen- oder Spaltenwerte gilt (oder für beide Werte), ist abhängig von der verwendeten Normalisierungsmethode. Jeder Zeilen- und Spaltenpunkt trägt zur Trägheit bei. Zeilen- und Spaltenpunkte, die beträchtlich zur Trägheit einer Dimension beitragen, sind für diese Dimension wichtig. Der Beitrag eines Punkts zur Trägheit einer Dimension entspricht der gewichteten quadrierten Distanz vom projizierten Punkt auf den Ursprung, dividiert durch die Trägheit der Dimension.

Die Diagnose, mit der die Beiträge der Punkte gemessen wird, ist eine wichtige Hilfe bei der Interpretation einer Korrespondenzanalysenlösung. Dominante Punkte in der Lösung sind leicht erkennbar. Beispiel: Die Angestellten mit und ohne Erfahrung tragen 84 % der Trägheit bei und bilden somit dominante Punkte in der ersten Dimension. Keiner der Spaltenpunkte trägt 65 % der Trägheit allein in der ersten Dimension bei.

Der Beitrag eines Punkts zur Trägheit der Dimensionen ist abhängig von der Masse und von der Distanz zum Ursprung. Punkte, die eine große Entfernung zum Ursprung und eine große Masse aufweisen, tragen den größten Teil zur Trägheit der Dimension bei. Zusatzpunkte spielen keine Rolle bei der Definition der Lösung und tragen somit auch nicht zur Trägheit der Dimensionen bei.

Neben dem Beitrag der Punkte zur Trägheit pro Dimension können Sie den Beitrag der Dimensionen zur Trägheit pro Punkt untersuchen. Um zu ermitteln, wie die Trägheit eines Punkts über die Dimensionen gestreut wird, berechnen Sie den Prozentsatz der Punktträgheit, der von den einzelnen Dimensionen beigetragen wird. Beachten Sie, dass die Beiträge der Dimensionen zu den Werten für die Punktträgheit zusammengerechnet nicht immer 1 ergeben. In einem reduzierten Raum wird die Trägheit aus den höheren Dimensionen nicht dargestellt. Wenn Sie die maximal zulässigen Dimensionen verwenden, werden auch die fehlenden Trägheitswerte sichtbar.

Die ersten beiden Dimensionen tragen die gesamte Trägheit für Angestellte mit und ohne Erfahrung bei sowie nahezu die gesamte Trägheit für Junior Manager und Sekretariatsangehörige. Bei Senior Managern werden 11 % der Trägheit nicht durch die ersten beiden Dimensionen beigetragen. Zwei Dimensionen tragen einen sehr großen Teil der Trägheiten für die Zeilenpunkte bei.

Ähnliche Ergebnisse sind bei den Spaltenpunkten zu verzeichnen. Bei jedem aktiven Spaltenpunkt tragen zwei Dimensionen mindestens 98 % der Trägheit bei. Die dritte Dimension liefert nur einen sehr kleinen Beitrag zu diesen Punkten.

Permutationen der Korrespondenztabelle

Unter Umständen sollten die Kategorien in den Zeilen und Spalten sortiert werden. Sie vermuten beispielsweise, dass die Kategorien einer Variablen eine bestimmte Reihenfolge aufweisen, kennen diese Reihenfolge jedoch nicht genau. Dieses Ordnungsproblem tritt in vielen Fachgebieten auf: Problem der reihenweisen Anordnung bei der Archäologie, Ordinationsproblem bei der Pflanzensoziologie sowie das Skalogrammproblem nach Guttman in den Gesellschaftswissenschaften. Zum Sortieren können die Zeilen- und Spaltenwerte als Sortiervariablen herangezogen werden. Wenn Zeilen- und Spaltenwerte in p Dimensionen vorliegen, können p Permutationstabellen erstellt werden. Ist der erste Singulärwert groß, zeigt die erste Tabelle eine bestimmte Struktur, bei der sich die relativen Häufigkeiten, die größer sind als erwartet, nahe an der "Diagonalen" befinden.

Die nachstehende Tabelle zeigt die Permutation der Korrespondenztabelle in der ersten Dimension. Wenn Sie die Zeilenwerte für Dimension 1 betrachten, wird die folgende Rangfolge von unten nach oben ersichtlich: Angestellte mit Erfahrung (–0,728), Sekretariatsangehörige (–0,385), Senior Manager (–0,126), Angestellte ohne Erfahrung (0,446), Junior Manager (0,495). Aus den Spaltenwerten für Dimension 1 geht die Rangliste "Nichtraucher", "Leicht", "Mittel" und "Stark" hervor. Diese Ranglisten werden beim Sortieren der Zeilen und Spalten in der Tabelle berücksichtigt.

Abbildung 12-13
Permutation der Korrespondenztabelle

	Rauchgewohnheit				
Berufsgruppe	Nichtraucher	Leicht	Mittel	Stark	Aktiver Rand
Angestellter mit Erfahrung	25	10	12	4	51
Sekretariat	10	6	7	2	25
Senior Manager	4	2	3	2	11
Angestellter ohne Erfahrung	18	24	33	13	88
Junior Manager	4	3	7	4	18
Aktiver Rand	61	45	62	25	193

Konfidenzstatistiken

Unter der Annahme, dass eine Häufigkeitstabelle analysiert werden soll und dass die Daten eine Zufallsstichprobe aus einer unbekannten Gesamtheit umfassen, folgen die Zellenhäufigkeiten einer multinomialen Verteilung. Ausgehend hiervon können die Standardabweichungen und die Korrelationen der Singulärwerte, der Zeilenwerte und der Spaltenwerte berechnet werden.

In einer eindimensionalen Korrespondenzanalyselösung können Sie je ein Konfidenzintervall für die einzelnen Werte in der Gesamtheit berechnen. Bei einer großen Standardabweichung bestehen in der Korrespondenzanalyse große Unsicherheiten hinsichtlich der Position, an der sich der Punkt in der Gesamtheit befindet. Bei einer kleinen Standardabweichung läßt sich in der Korrespondenzanalyse dagegen relativ sicher sagen, dass dieser Punkt sehr nahe an dem Punkt aus der Lösung liegt.

Bei einer mehrdimensionalen Lösung mit hoher Korrelation zwischen den Dimensionen ist es unter Umständen nicht möglich, einen Punkt mit angemessener Sicherheit in der richtigen Dimension aufzufinden. In solchen Fällen sind multivariate Konfidenzintervalle mithilfe der Varianz-Kovarianz-Matrix zu berechnen, die in eine Datei geschrieben werden kann.

Die Konfidenzstatistiken für die Zeilen- und Spaltenwerte werden gezeigt. Die Standardabweichungen für die beiden Manager-Kategorien sind größer als die anderen Werte; dies liegt wahrscheinlich an der relativ geringen Anzahl an Personen in diesen Kategorien. Auch die Standardabweichung für starke Raucher ist aus demselben Grund größer. Wenn Sie die Korrelationen zwischen den Dimensionen für die Werte betrachten, wird ersichtlich, dass die Korrelationen für die Zeilen- und Spaltenwerte im allgemeinen klein sind, außer bei den Angestellten ohne Erfahrung; hier liegt ein Wert von 0,611 für die Korrelation vor.

Abbildung 12-14
Konfidenzstatistiken für Zeilenwerte

	Standardabweichung in Dimension		Korrelation
Berufsgruppe	1	2	1-2
Senior Manager	,614	,917	,101
Junior Manager	,461	,511	,007
Angestellter mit Erfahrung	,110	,157	,107
Angestellter ohne Erfahrung	,118	,124	,611
Sekretariat	,158	,153	-,360

Abbildung 12-15
Konfidenzstatistiken für Spaltenwerte

	Standardabweichung in Dimension		Korrelation
Rauchgewohnheit	1	2	1-2
Nichtraucher	,118	,145	,402
Leicht	,281	,292	,054
Mittel	,179	,332	,020
Stark	,361	,441	-,155

Zusatzprofile

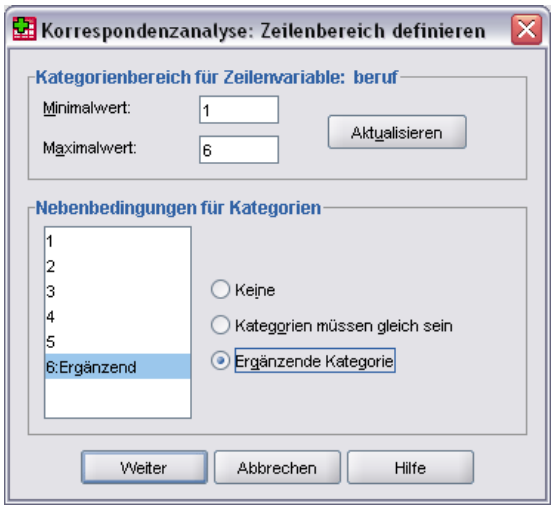
Bei der Korrespondenzanalyse können zusätzliche Kategorien im Raum repräsentiert werden, die die Beziehung zwischen den aktiven Kategorien beschreiben. Mit einem Zusatzprofil wird lediglich ein Profil über die Kategorien der Zeilen- oder der Spaltenvariable definiert; dieses Profil wirkt sich in keinster Weise auf die Analyse aus. Die Datendatei enthält eine Zusatzzeile und zwei Zusatzspalten.

Der Landesdurchschnitt in den Kategorien für die Rauchgewohnheiten bildet ein zusätzliches Zeilenprofil. Die beiden Zusatzspalten definieren zwei Spaltenprofile über die Berufsgruppen. Mit den Zusatzprofilen wird ein Punkt definiert, der entweder im Zeilenraum oder im Spaltenraum

liegt. Die Zeilen und Spalten werden getrennt voneinander betrachtet; verwenden Sie daher die Prinzipal-Normalisierung.

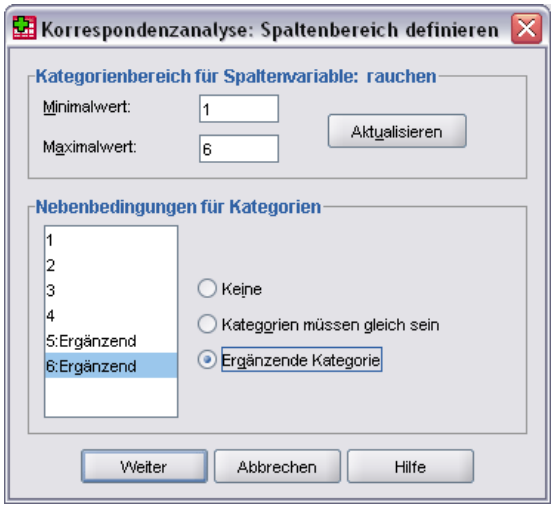
Durchführung der Analyse

Abbildung 12-16
Dialogfeld "Zeilenbereich definieren"



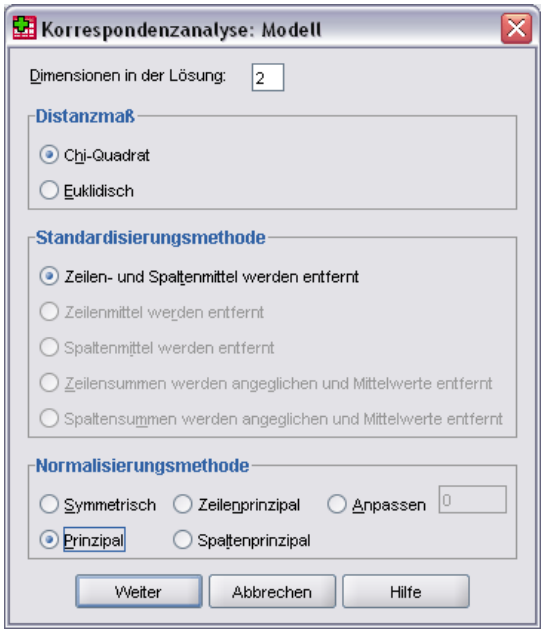
- Um die Zusatzkategorien hinzuzufügen und eine Prinzipal-Normalisierungslösung zu erstellen, öffnen Sie erneut das Dialogfeld "Korrespondenzanalyse".
- Wählen Sie *staff* und klicken Sie auf Bereich definieren.
- Geben Sie 6 als Höchstwert ein und klicken Sie auf Aktualisieren.
- Wählen Sie in der Liste "Nebenbedingungen für Kategorien" den Wert 6 aus und wählen Sie die Option Ergänzende Kategorie aus.
- Klicken Sie auf Weiter.
- Wählen Sie *Rauchen* und klicken Sie im Dialogfeld "Korrespondenzanalyse" auf Bereich definieren.

Abbildung 12-17
Dialogfeld "Spaltenbereich definieren"



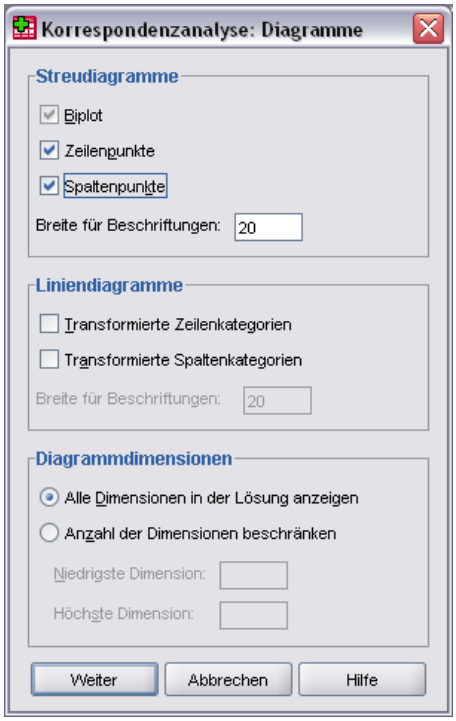
- Geben Sie 6 als Höchstwert ein und klicken Sie auf Aktualisieren.
- Wählen Sie in der Liste "Nebenbedingungen für Kategorien" den Wert 5 aus und wählen Sie die Option Ergänzende Kategorie aus.
- Wählen Sie in der Liste "Nebenbedingungen für Kategorien" den Wert 6 aus und wählen Sie die Option Ergänzende Kategorie aus.
- Klicken Sie auf Weiter.
- Klicken Sie im Dialogfeld "Korrespondenzanalyse" auf Modell.

Abbildung 12-18
Dialogfeld "Modell"



- Wählen Sie Prinzipal als Normalisierungsmethode.
- Klicken Sie auf Weiter.
- Klicken Sie im Dialogfeld "Korrespondenzanalyse" auf Diagramme.

Abbildung 12-19
Dialogfeld "Diagramme"

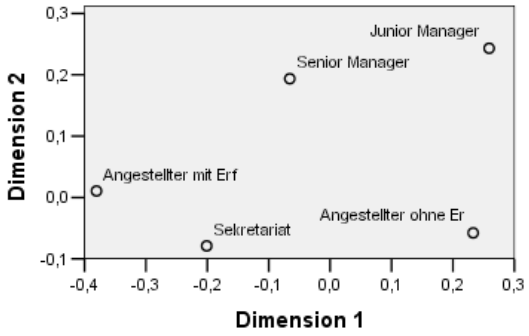


- Wählen Sie in der Gruppe "Streudiagramme" die Optionen Zeilenpunkte und Spaltenpunkte.
- Klicken Sie auf Weiter.
- Klicken Sie im Dialogfeld "Korrespondenzanalyse" auf OK.

Das Diagramm für die Zeilenpunkte zeigt die ersten beiden Dimensionen für die Zeilenpunkte und dazu den Zusatzpunkt für *National Average*. *National Average* liegt weit entfernt von Ursprung; dies bedeutet, dass die Stichprobe im Hinblick auf die Rauchgewohnheiten nicht als repräsentativ für das Land gelten kann. Die Gruppen "Sekretariat" und "Angestellter mit Erfahrung" befinden sich nahe am Landesdurchschnitt, im Gegensatz zur Gruppe "Junior

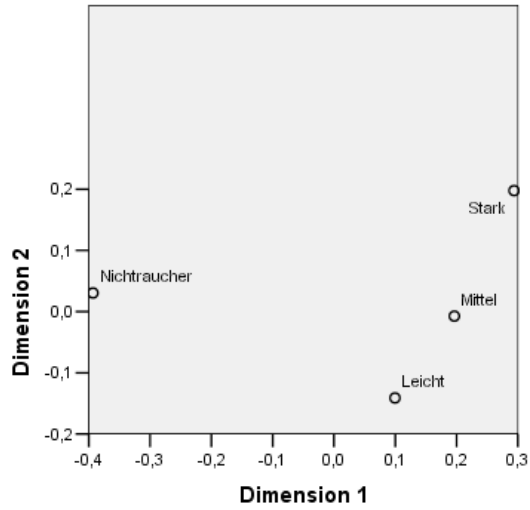
Manager". Sekretariatsangehörige sowie Angestellte mit Erfahrung zeigen demnach ähnliche Rauchgewohnheiten wie der Landesdurchschnitt, Junior Manager dagegen nicht.

Abbildung 12-20
Zeilenpunkte (Prinzipal-Normalisierung)



Das Diagramm für die Spaltenpunkte zeigt den Spaltenraum mit den beiden Zusatzpunkten für den Alkoholkonsum. *Alcohol* liegt nahe am Ursprung und zeigt so eine enge Korrespondenz zwischen dem Alkoholprofil und dem durchschnittlichen Spaltenprofil. *No Alcohol* weicht dagegen vom durchschnittlichen Spaltenprofil ab, was durch die große Distanz zum Ursprung verdeutlicht wird. Der nächstgelegene Punkt zu *No Alcohol* ist *Leicht*. Das Profil der leichten Raucher weist die größten Ähnlichkeiten mit den Abstinenzlern auf. Unter den Rauchern liegt der Punkt *Mittel* am nächsten und der Punkt *Schwer* am weitesten entfernt. Beim leichten bis schweren Rauchen besteht somit eine Progression im Hinblick auf die Ähnlichkeit mit der Abstinenz. Durch den relativ hohen Anteil der Sekretariatsangehörigen in der Gruppe *No Alcohol* ist allerdings keine enge Korrespondenz zu den Raucher-Kategorien gegeben.

Abbildung 12-21
Spaltenpunkte (Prinzipal-Normalisierung)



Beispiel: Wahrnehmung von Kaffeemarken

Das vorangegangene Beispiel umfasste eine kleine Tabelle mit hypothetischen Daten. Im "wirklichen Leben" fallen häufig deutlich größere Tabellen an. In diesem Beispiel verwenden Sie Daten zum wahrgenommenen Image von sechs Eiskaffeemarken (Kennedy, Riquier, als auch Sharp, 1996). Dieses Daten-Set finden Sie in *coffee.sav*. Für weitere Informationen siehe Beispieldateien in Anhang A auf S. 326.

Bei den 23 Attributen des Eiskaffee-Image sollten die Teilnehmer jeweils alle Marken auswählen, die durch dieses Attribut beschrieben werden. Die sechs Marken werden als *AA*, *BB*, *CC*, *DD*, *EE* und *FF* bezeichnet, um Vertraulichkeit zu gewährleisten.

Tabelle 12-1
Eiskaffeeattribute

Image-Attribut	Label	Image-Attribut	Label
Gutes Heilmittel gegen Kater	*Heilmittel*	Dick machende Marke	*macht dick*
Fettarme/kalorienarme Marke	*fettarm*	Ansprechend für Männer	*Männer*
Marke für Kinder	*Kinder*	Südaustralische Marke	*Süd Australien*
Marke für die "Arbeiterklasse"	*Arbeit*	Traditionelle/alteingesessene Marke	*traditionell*
Reichhaltige/süße Marke	*süss*	Marke mit Premium-Qualität	*premium*
Unpopuläre Marke	*unpopulär*	Gesunde Marke	*gesund*
Marke für dicke/häßliche Menschen	*hässlich*	Marke mit hohem Koffeingehalt	*Koffein*
Sehr erfrischend	*frisch*	Neue Marke	*neu*
Marke für Yuppies	*Yuppies*	Marke für attraktive Menschen	*attraktiv*
Nahrhafte Marke	*nährstoffreich*	Schwere Marke	*schwer*
Marke für Frauen	*Frauen*	Populäre Marke	*populär*
Gering vertretene Marke	*gering*		

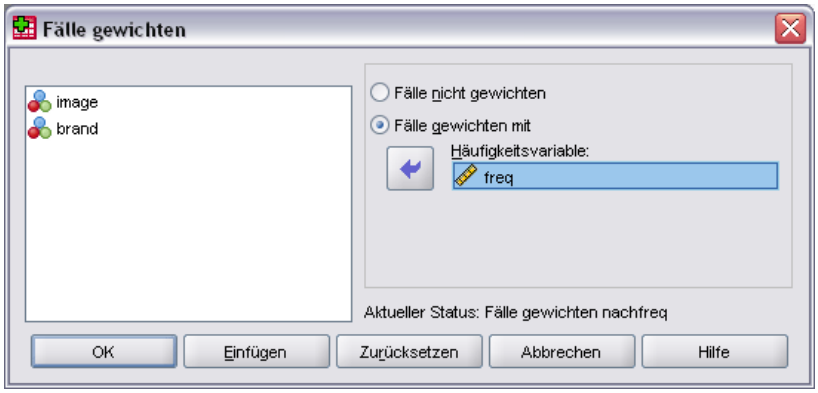
Zu Beginn konzentrieren Sie sich darauf, wie die Attribute miteinander zusammenhängen und in welchem Zusammenhang die Marken stehen. Mit der Prinzipal-Normalisierung wird die Gesamtträgheit einmal über die Zeilen und einmal über die Spalten gestreut. So steht zwar die Biplot-Interpretation nicht mehr zur Verfügung, aber die Distanzen zwischen den Kategorien der einzelnen Variablen können untersucht werden.

Durchführung der Analyse

▶ Gemäß dem Aufbau der Daten müssen die Fälle nach der Variablen *Häufigke* gewichtet werden. Wählen Sie hierzu die folgenden Befehle aus den Menüs aus:

Daten
  Fälle gewichten...

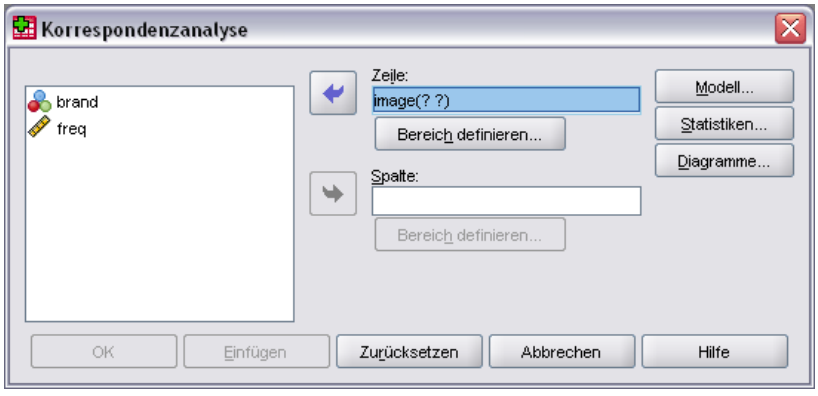
Abbildung 12-22
Dialogfeld "Fälle gewichten"



- Fälle gewichten nach *freq*.
- Klicken Sie auf OK.
- Zum Erstellen einer anfänglichen Lösung in fünf Dimensionen mit Prinzipal-Normalisierung wählen Sie die folgenden Befehle aus den Menüs aus:

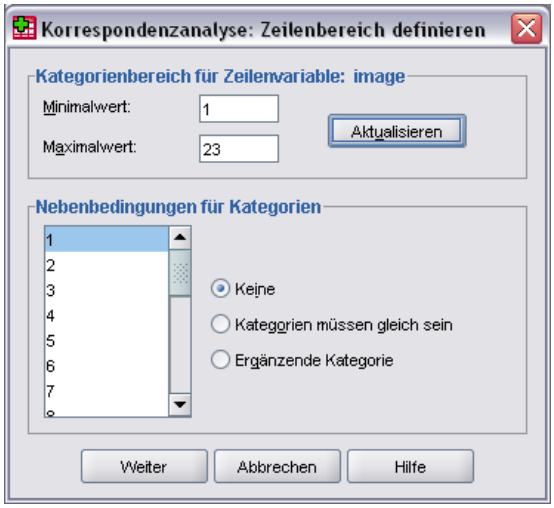
Analysieren
  Dimensionsreduktion
    Korrespondenzanalyse...

Abbildung 12-23
Dialogfeld "Korrespondenzanalyse"



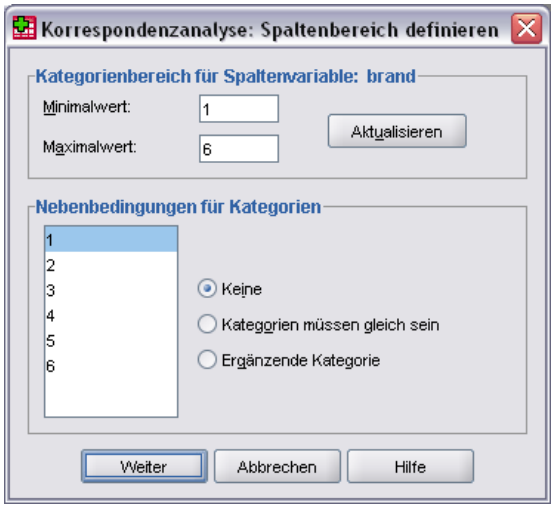
- Wählen Sie *Image* als Zeilenvariable aus.
- Klicken Sie auf Bereich definieren.

Abbildung 12-24
Dialogfeld "Zeilenbereich definieren"



- Geben Sie 1 als Mindestwert ein.
- Geben Sie 23 als Höchstwert ein.
- Klicken Sie auf Aktualisieren.
- Klicken Sie auf Weiter.
- Wählen Sie *Marke* als Spaltenvariable aus.
- Klicken Sie im Dialogfeld "Korrespondenzanalyse" auf Bereich definieren.

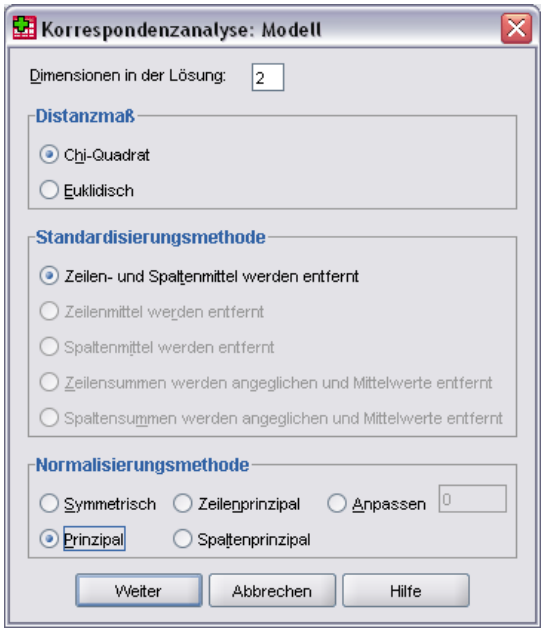
Abbildung 12-25
Dialogfeld "Spaltenbereich definieren"



- Geben Sie 1 als Mindestwert ein.
- Geben Sie 6 als Höchstwert ein.

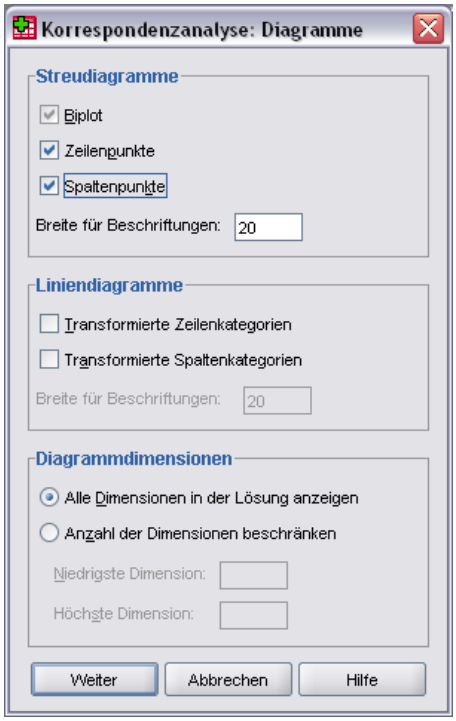
- Klicken Sie auf Aktualisieren.
- Klicken Sie auf Weiter.
- Klicken Sie im Dialogfeld “Korrespondenzanalyse” auf Modell.

Abbildung 12-26
Dialogfeld “Modell”



- Wählen Sie Prinzipal als Normalisierungsmethode.
- Klicken Sie auf Weiter.
- Klicken Sie im Dialogfeld “Korrespondenzanalyse” auf Diagramme.

Abbildung 12-27
Dialogfeld "Diagramme"



- Wählen Sie in der Gruppe "Streudiagramme" die Optionen Zeilenpunkte und Spaltenpunkte.
- Klicken Sie auf Weiter.
- Klicken Sie im Dialogfeld "Korrespondenzanalyse" auf OK.

Dimensionen

Die Trägheit pro Dimension zeigt die Zerlegung der Gesamtträgheit auf die einzelnen Dimensionen. Zwei Dimensionen machen 83 % der Gesamtträgheit aus. Bei einer dritten Dimension würden nur weitere 8,6 % der Trägheit berücksichtigt. Sie entscheiden sich daher für eine zweidimensionale Darstellung.

Abbildung 12-28
Trägheit pro Dimension

Dimension	Singulärwert	Auswertung für Trägheit	Chi-Quadrat	Sig.	Anteil der Trägheit		Singulärwert für Konfidenz	
					Bedingen	Kumuliert	Standard abweichung	Korrelation 2
1	,711	,506			,629	,629	,009	,132
2	,399	,159			,198	,827	,014	
3	,263	,069			,086	,913		
4	,234	,055			,068	,982		
5	,121	,015			,018	1,000		
Gesamtauswertung		,804	3746,97	,000[a]	1,000	1,000		

a. 110 Freiheitsgrade

Beiträge

Die Übersicht der Zeilenpunkte zeigt die Beiträge der Zeilenpunkte zur Trägheit der Dimensionen sowie den Beitrag der Dimensionen zur Trägheit der Zeilenpunkte. Wenn alle Punkte gleichmäßig zur Trägheit beitrügen, wäre der Beitrag gleich 0,043. Die Attribute *gesund* und *fettarm* tragen beide einen beträchtlichen Anteil zur Trägheit der ersten Dimension bei. Die Attribute *Männer* und *stark* liefern die größten Beiträge zur Trägheit der zweiten Dimension. Sowohl *hässlich* als auch *frisch* bringen nur sehr geringe Beiträge zu den Dimensionen.

Abbildung 12-29
Beiträge der Attribute

		Wert in Dimension			Beitrag				
					des Punktes an der Trägheit der Dimension		der Dimension an der Trägheit des Punktes		
image	Masse	1	2	Übersicht über Trägheit	1	2	1	2	Gesamtübersicht
macht dick	,080	-,514	-,265	,033	,042	,035	,652	,173	,825
Männer	,051	-,852	,825	,072	,073	,219	,512	,480	,992
Süd Australien	,057	-,303	-,350	,046	,010	,044	,114	,152	,266
traditionell	,040	-,703	-,532	,043	,039	,071	,454	,260	,715
premium	,042	-,444	-,582	,028	,016	,090	,296	,509	,805
gesund	,053	1,200	,174	,081	,152	,010	,953	,020	,973
Koffein	,047	-,452	,124	,014	,019	,005	,702	,053	,755
neu	,047	,960	,147	,048	,086	,006	,893	,021	,914
attraktiv	,041	,657	-,056	,019	,035	,001	,911	,007	,918
schwer	,039	-,850	1,002	,070	,056	,246	,404	,560	,964
populär	,060	-,697	-,042	,038	,058	,001	,771	,003	,774
Heilmittel	,026	-,389	,266	,009	,008	,011	,446	,209	,655
fettarm	,052	1,305	,196	,094	,175	,013	,941	,021	,962
Kinder	,024	-,352	-,513	,017	,006	,041	,179	,380	,559
Arbeit	,045	-,785	,477	,040	,055	,064	,693	,255	,948
süss	,038	-,519	-,683	,048	,020	,112	,212	,368	,580
unpopulär	,024	,489	,186	,010	,011	,005	,585	,085	,670
hässlich	,030	,006	-,109	,003	,000	,002	,000	,131	,131
frisch	,036	-,096	-,100	,002	,001	,002	,196	,214	,410
Yuppies	,034	,380	-,301	,012	,010	,019	,392	,246	,637
nährstoffreich	,040	,722	,055	,022	,041	,001	,946	,006	,951
Frauen	,054	,758	-,063	,032	,062	,001	,965	,007	,972
gering	,040	,579	,063	,023	,027	,001	,593	,007	,600
Aktiver Gesamtwert	1,000			,804	1,000	1,000			

Zwei Dimensionen tragen einen Großteil der Trägheit bei den meisten Zeilenpunkten bei. Die großen Anteile der ersten Dimension an *gesund*, *neu*, *attraktiv*, *fettarm*, *nährstoffreich* und *Frauen* zeigen, dass diese Punkte in einer Dimension sehr stark repräsentiert sind. Die höheren Dimensionen tragen folglich nur wenig zur Trägheit dieser Punkte bei, die sehr nahe an der horizontalen Achse liegen. Die zweite Dimension liefert die größten Beiträge für *Männer*, *premium* und *stark*. Beide Dimensionen haben einen sehr geringen Anteil an der Trägheit für *Süd Australien* und *hässlich*; diese Punkte sind daher unterrepräsentiert.

Die Übersicht der Spaltenpunkte zeigt die Beiträge im Zusammenhang mit den Spaltenpunkten. Die Marken *CC* und *DD* tragen am stärksten zur ersten Dimension bei, die Marken *EE* und *FF* erklären einen Großteil der Trägheit für die zweite Dimension. Sowohl *AA* als auch *BB* bringen nur sehr geringe Beiträge zu den Dimensionen.

Abbildung 12-30
Beiträge der Marken

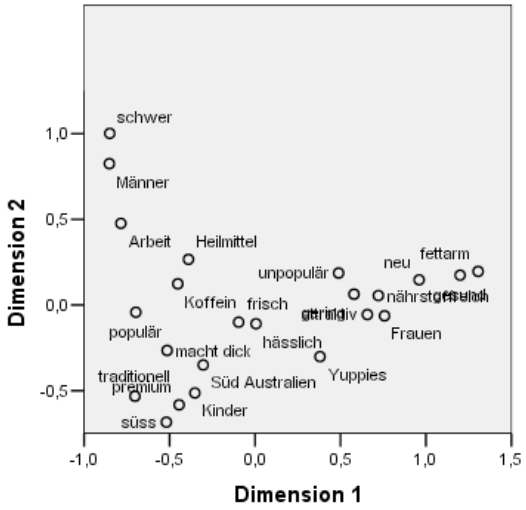
		Wert in Dimension			Beitrag				
					des Punktes an der Trägheit der Dimension		der Dimension an der Trägheit des Punktes		
marke	Masse	1	2	Übersicht über Trägheit	1	2	1	2	Gesamtübersicht
AA	,217	-,659	,046	,127	,187	,003	,744	,004	,748
BB	,131	-,284	-,404	,078	,021	,134	,135	,272	,407
CC	,185	,996	,076	,193	,362	,007	,951	,006	,957
DD	,162	,915	,101	,146	,267	,010	,928	,011	,939
EE	,152	-,651	,706	,153	,127	,477	,420	,494	,914
FF	,153	-,343	-,618	,107	,036	,369	,169	,550	,718
Aktiver Gesamtwert	1,000			,804	1,000	1,000			

In beiden Dimensionen sind alle Marken gut repräsentiert, mit Ausnahme von *BB*. *CC* und *DD* sind gut in einer Dimension vertreten. Die zweite Dimension liefert die größten Beiträge für *EE* und *FF*. Beachten Sie, dass die Marke *AA* in der ersten Dimension zwar gut repräsentiert ist, jedoch keinen sehr hohen Beitrag zu dieser Dimension liefert.

Diagramme

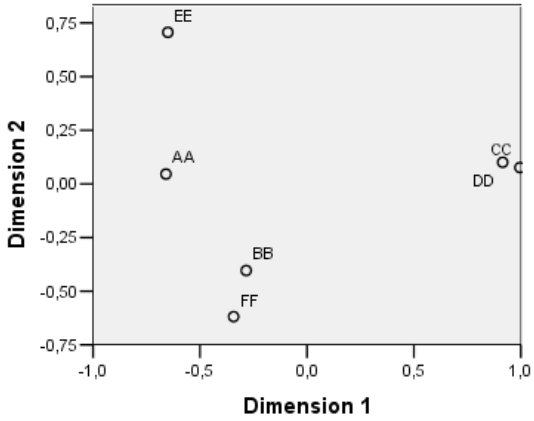
Die Zeilenpunktdiagramme zeigen, dass die Attribute *frisch* und *hässlich* sehr nahe am Ursprung liegen; dies bedeutet, dass diese Attribute nur wenig vom durchschnittlichen Zeilenprofil abweichen. Drei allgemeine Klassifizierungen werden deutlich. Im oberen linken Bereich finden sich die Attribute *stark*, *Männer* und *Arbeit*, jeweils mit großen Ähnlichkeiten. Der untere linke Bereich enthält *süss*, *macht dick*, *Kinder* und *premium*. Die Attribute *gesund*, *fettarm*, *nährstoffreich* und *neu* drängen sich dagegen im rechten Teil des Diagramms.

Abbildung 12-31
Diagramm der Image-Attribute (Prinzipal-Normalisierung)



Im Spaltenpunktdiagramm liegen alle Marken weit enfernt vom Ursprung; keine der Marken ähnelt also dem Gesamt-Zentroid. Die Marken *CC* und *DD* liegen gemeinsam im rechten Bereich, die Marken *BB* und *FF* treffen sich dagegen im unteren Teil des Diagramms. Die Marken *AA* und *EE* weisen keine Ähnlichkeiten mit den anderen Marken auf.

Abbildung 12-32
Diagramm der Marken (Prinzipal-Normalisierung)

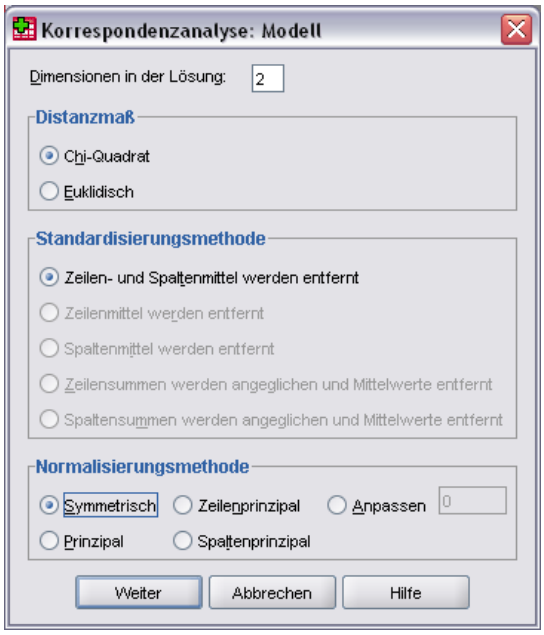


Symmetrische Normalisierung

Wie hängen die Marken mit den Image-Attributen zusammen? Die Prinzipal-Normalisierung ist für diese Beziehungen nicht geeignet. Um den Schwerpunkt auf den Zusammenhang untereinander zu legen, verwenden Sie die symmetrische Normalisierung. Die Trägheit wird nicht zweimal gestreut (wie bei der Prinzipal-Normalisierung), sondern gleichmäßig auf die Zeilen und Spalten aufgeteilt. Die Distanzen zwischen den Kategorien einer bestimmten Variablen können

nicht interpretiert werden; die Distanzen zwischen den Kategorien verschiedener Variablen sind dagegen sehr wohl aussagekräftig.

Abbildung 12-33
Dialogfeld "Modell"

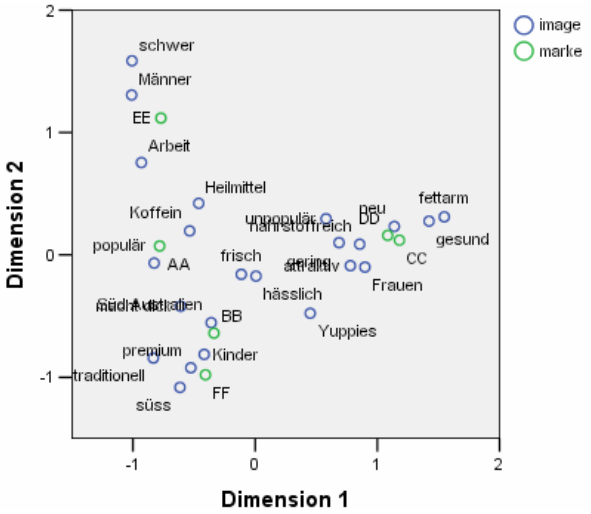


- Zum Anlegen der nachstehenden Lösung mit symmetrischer Normalisierung öffnen Sie erneut das Dialogfeld "Korrespondenzanalyse" und klicken Sie auf Modell.
- Wählen Sie Symmetrisch als Normalisierungsmethode.
- Klicken Sie auf Weiter.
- Klicken Sie im Dialogfeld "Korrespondenzanalyse" auf OK.

Aus dem oberen linken Bereich des Biplot geht die Marke *EE* als einzige starke Marke für die "Arbeiterklasse" hervor, die zudem für Männer ansprechend ist. Die Marke *AA* ist die populärste Marke und gilt außerdem als Marke mit dem höchsten Koffeingehalt. Zu den süßen, dick

machenden Marken zählen die Marken *BB* und *FF*. Die Marken *CC* und *DD* werden zwar als neu und gesund empfunden, sind jedoch auch die unpopulärsten Marken.

Abbildung 12-34
Biplot der Marken und Attribute (symmetrische Normalisierung)



Zur weiteren Interpretation können Sie eine Linie durch den Ursprung und die beiden Image-Attribute *Männer* und *Yuppies* ziehen und dann die Marken auf diese Linie projizieren. Die beiden Attribute stehen einander gegenüber. Dies bedeutet, dass das Assoziationsmuster der Marken für *Männer* dem Muster für *Yuppies* entgegen läuft. Männer werden also am häufigsten mit der Marke *EE* und am wenigsten häufig mit der Marke *CC* verbunden, Yuppies dagegen am häufigsten mit der Marke *CC* und am wenigsten häufig mit der Marke *EE*.

Beispiel: Flugmeilen zwischen Städten

Die Korrespondenzanalyse ist nicht auf bloße Häufigkeitstabellen beschränkt. Die Einträge können eine beliebige positive Messzahl der Korrespondenz darstellen. In diesem Beispiel verwenden Sie die Flugmeilen zwischen zehn Städten in den USA. Dieses Daten-Set finden Sie in *flying.sav*. Für weitere Informationen siehe Beispieldateien in Anhang A auf S. 326.

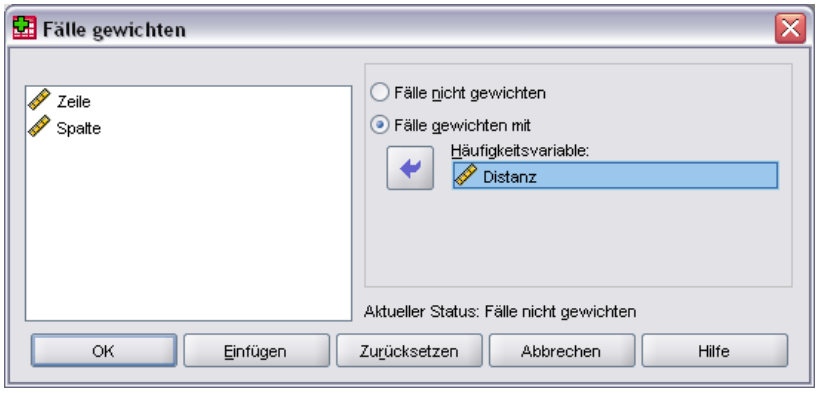
Tabelle 12-2
Städtenamen

Ort	Label	Ort	Label
Atlanta	*Atl*	Miami	*Mia*
Chicago	*Chi*	New York	*NY*
Denver	*Den*	San Francisco	*SF*
Houston	*Hou*	Seattle	*Sea*
Los Angeles	*LA*	Washington, DC	*DC*

- Zum Anzeigen der Flugmeilen gewichten Sie die Fälle zunächst nach der Variablen *Distanz*. Wählen Sie die folgenden Befehle aus den Menüs aus:

  Daten
  Fälle gewichten...

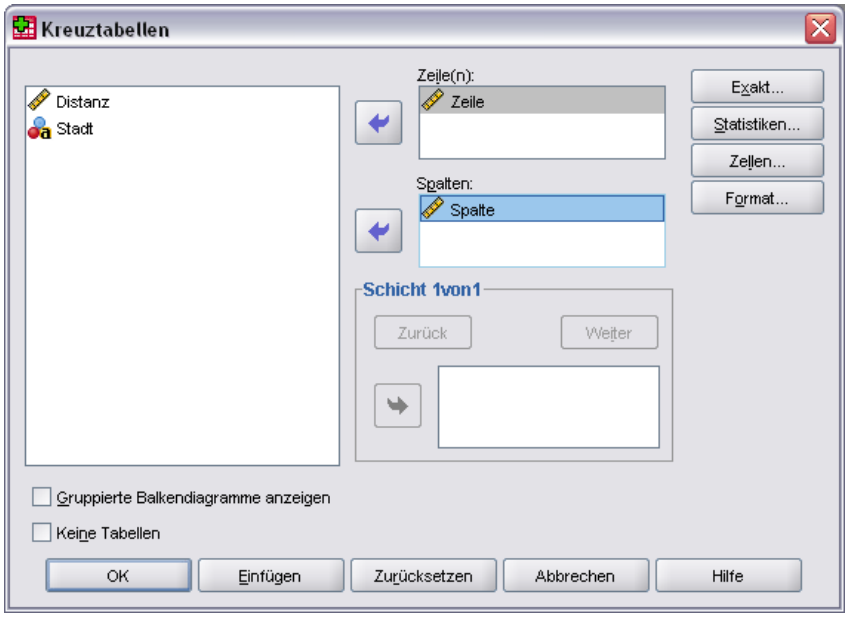
Abbildung 12-35
Dialogfeld "Fälle gewichten"



- Fälle gewichten nach *Distanz*.
- Klicken Sie auf OK.
- Zum Anzeigen der Meilen als Kreuztabelle wählen Sie die folgenden Befehle aus den Menüs aus:

  Analysieren
  Deskriptive Statistiken
  Kreuztabellen...

Abbildung 12-36
Dialogfeld "Kreuztabellen"



- Wählen Sie *Zeile* als Zeilenvariable aus.
- Wählen Sie *Spalte* als Spaltenvariable aus.
- Klicken Sie auf OK.

Die nachstehende Tabelle zeigt die Flugmeilen zwischen den Städten. Beachten Sie, dass nur eine einzige Variable für die Zeilen und die Spalten vorliegt und dass die Tabelle symmetrisch ist: Die Distanz von Los Angeles nach Miami ist identisch mit der Distanz von Miami noch Los Angeles. Darüber hinaus ist die Distanz einer Stadt zu sich selbst gleich 0. Der aktive Rand zeigt die Gesamtflugmeilen von einer beliebigen Stadt zu allen anderen Städten.

Abbildung 12-37
Flugmeilen zwischen 10 Städten in den USA

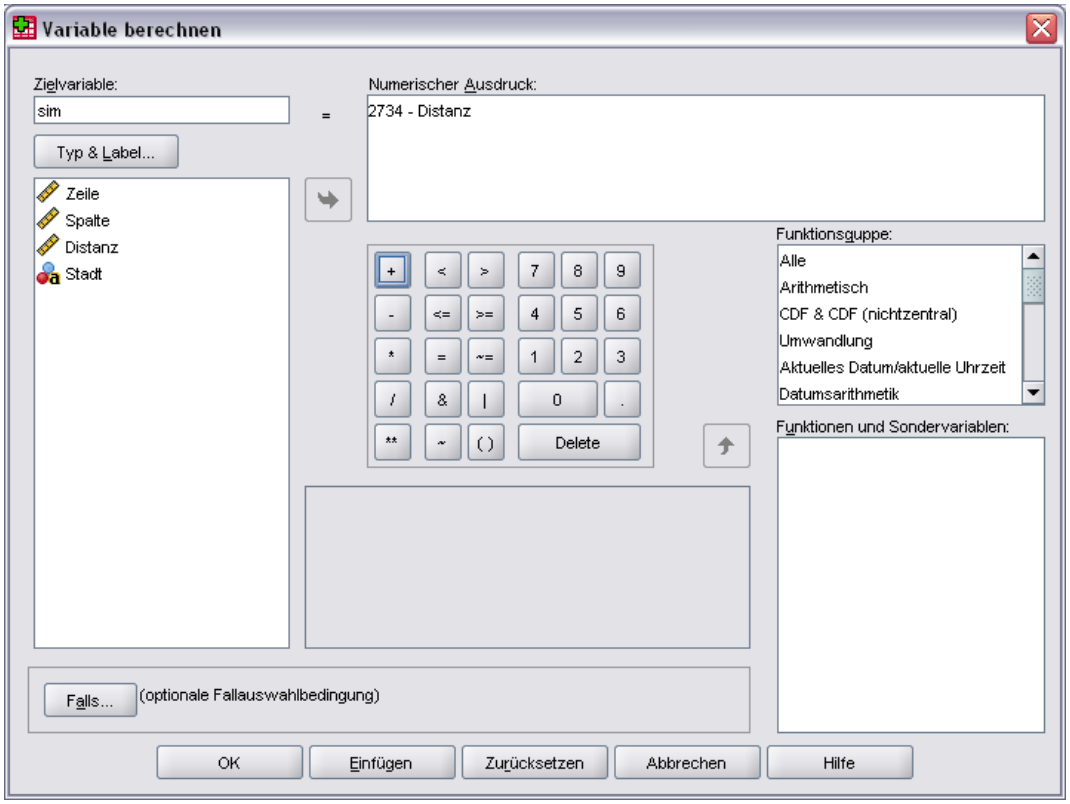
		Spalte										
		Atlanta	Chicago	Denver	Houston	Los Angeles	Miami	New York	San Francisco	Seattle	Washington DC	Gesamt
Zeile	Atl	0	587	1212	701	1936	604	748	2139	2182	543	10652
	Chi	587	0	920	940	1745	1188	713	1858	1737	597	10285
	Den	1212	920	0	879	831	1726	1631	949	1021	1494	10663
	Hou	701	940	879	0	1374	968	1420	1645	1891	1220	11038
	LA	1936	1745	831	1374	0	2339	2451	347	959	2300	14282
	Mia	604	1188	1726	968	2339	0	1092	2594	2734	923	14168
	NY	748	713	1631	1420	2451	1092	0	2571	2408	205	13239
	SF	2139	1858	949	1645	347	2594	2571	0	678	2442	15223
	Sea	2182	1737	1021	1891	959	2734	2408	678	0	2329	15939
	DC	543	597	1494	1220	2300	923	205	2442	2329	0	12053
Gesamt		10652	10285	10663	11038	14282	14168	13239	15223	15939	12053	127542

Im allgemeinen sind Distanzen Unähnlichkeiten; große Werte deuten auf große Differenzen zwischen den Kategorien hin. Für die Korrespondenzanalyse ist jedoch ein Zusammenhangsmaß erforderlich. Die Unähnlichkeiten müssen also in Ähnlichkeiten umgewandelt werden. In anderen Worten: Ein umfangreicher Tabelleneintrag muss einer kleinen Differenz zwischen den Kategorien entsprechen. Um die Unähnlichkeiten in Ähnlichkeiten umzuwandeln, subtrahieren Sie die einzelnen Tabelleneinträge jeweils vom größten Tabelleneintrag.

▶ Um die Ähnlichkeiten zu erstellen und in der neuen Variablen *Ähnlichk* zu speichern, wählen Sie die folgenden Befehle aus den Menüs aus:

Transformieren
  Variable berechnen...

Abbildung 12-38
Dialogfeld "Variable berechnen"

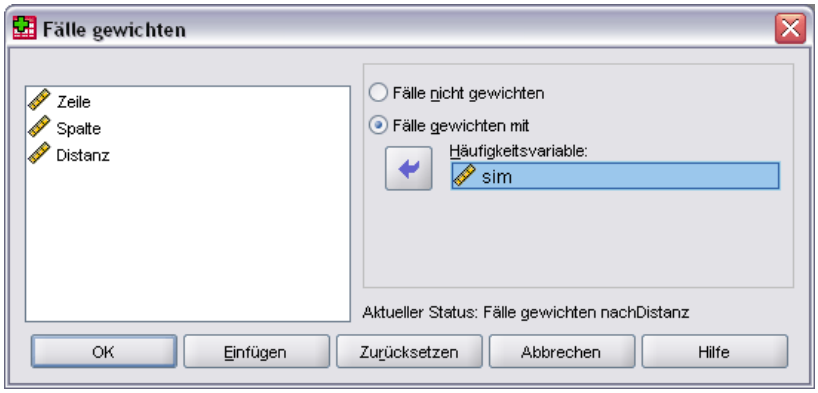


▶ Geben Sie sim als Zielvariable ein.

▶ Geben Sie 2734-Distanz als numerischen Ausdruck ein.

▶ Klicken Sie auf OK.

Abbildung 12-39
Dialogfeld "Fälle gewichten"

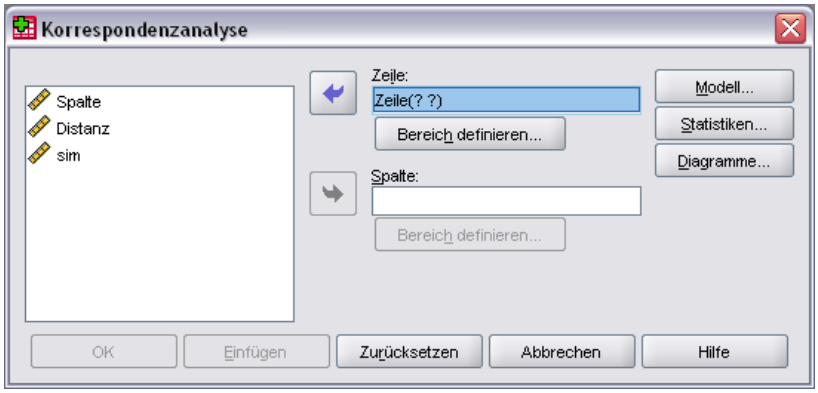


Gewichten Sie die Fälle nun nach dem Ähnlichkeitsmaß neu. Öffnen Sie dazu erneut das Dialogfeld "Fälle gewichten":

- Fälle gewichten nach *sim*.
- Klicken Sie auf OK.
- Zum Erstellen einer Korrespondenzanalyse für die Ähnlichkeiten wählen Sie schließlich die folgenden Befehle aus den Menüs aus:

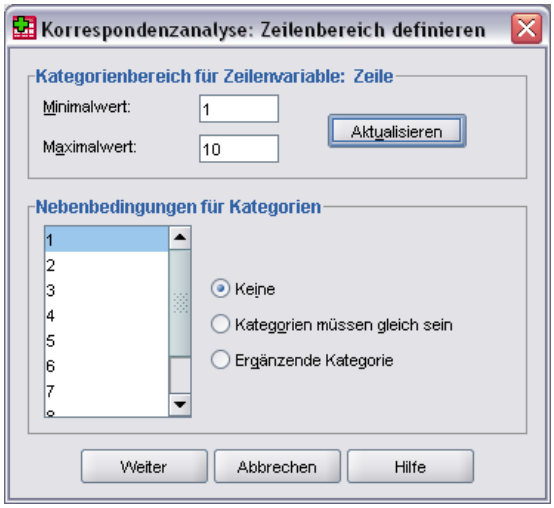
Analysieren
  Dimensionsreduktion
    Korrespondenzanalyse...

Abbildung 12-40
Dialogfeld "Korrespondenzanalyse"



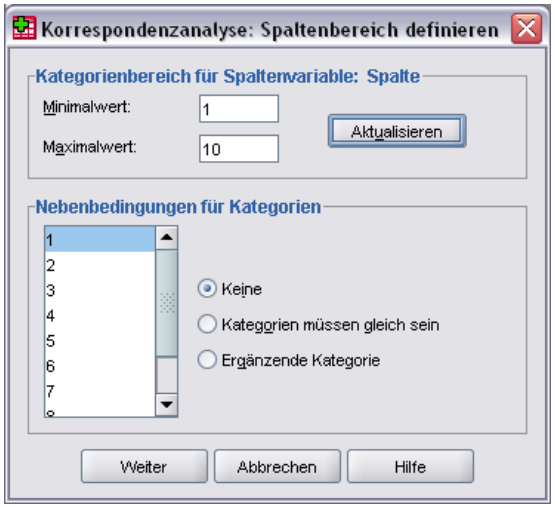
- Wählen Sie *Zeile* als Zeilenvariable aus.
- Klicken Sie auf Bereich definieren.

Abbildung 12-41
Dialogfeld "Zeilenbereich definieren"



- Geben Sie 1 als Mindestwert ein.
- Geben Sie 10 als Höchstwert ein.
- Klicken Sie auf Aktualisieren.
- Klicken Sie auf Weiter.
- Wählen Sie *Spalte* als Spaltenvariable aus.
- Klicken Sie im Dialogfeld "Korrespondenzanalyse" auf Bereich definieren.

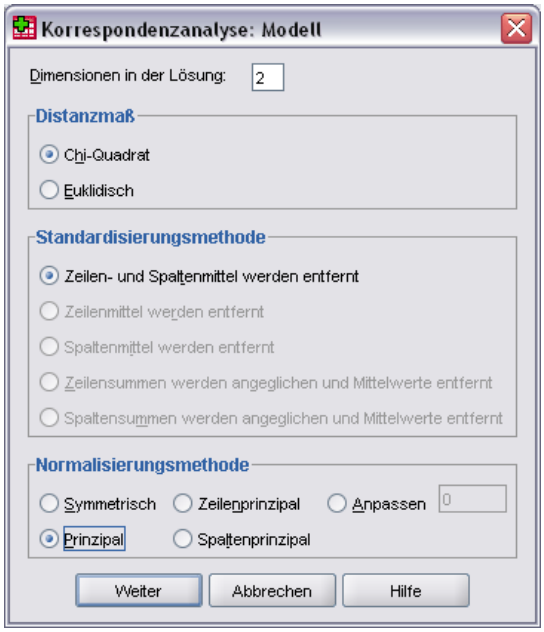
Abbildung 12-42
Dialogfeld "Spaltenbereich definieren"



- Geben Sie 1 als Mindestwert ein.
- Geben Sie 10 als Höchstwert ein.

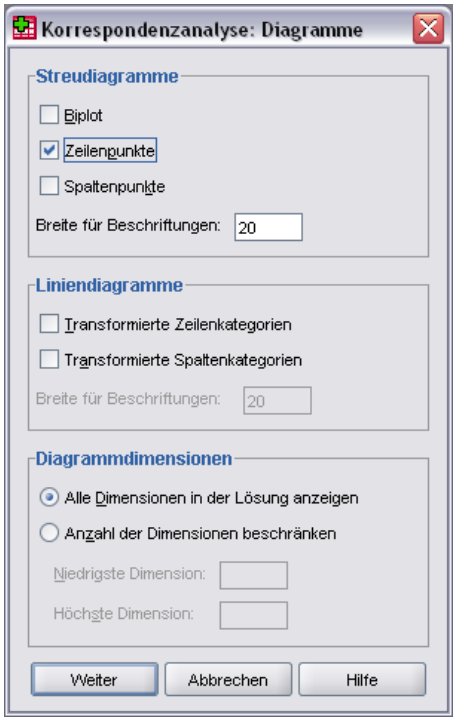
- Klicken Sie auf Aktualisieren.
- Klicken Sie auf Weiter.
- Klicken Sie im Dialogfeld “Korrespondenzanalyse” auf Modell.

Abbildung 12-43
Dialogfeld “Modell”



- Wählen Sie Prinzipal als Normalisierungsmethode.
- Klicken Sie auf Weiter.
- Klicken Sie im Dialogfeld “Korrespondenzanalyse” auf Diagramme.

Abbildung 12-44
Dialogfeld "Diagramme"



- Wählen Sie in der Gruppe "Streudiagramme" die Option Zeilenpunkte.
- Klicken Sie auf Weiter.
- Klicken Sie im Dialogfeld "Korrespondenzanalyse" auf OK.

Korrespondenztabelle.

Die neue Distanz 0 zwischen Seattle und Miami bedeutet, dass diese Städte am weitesten voneinander entfernt sind (geringste Ähnlichkeit); die Distanz 2529 zwischen New York und Washington, D.C., bedeutet dagegen, dass diese Städte einander am nächsten liegen (größte Ähnlichkeit).

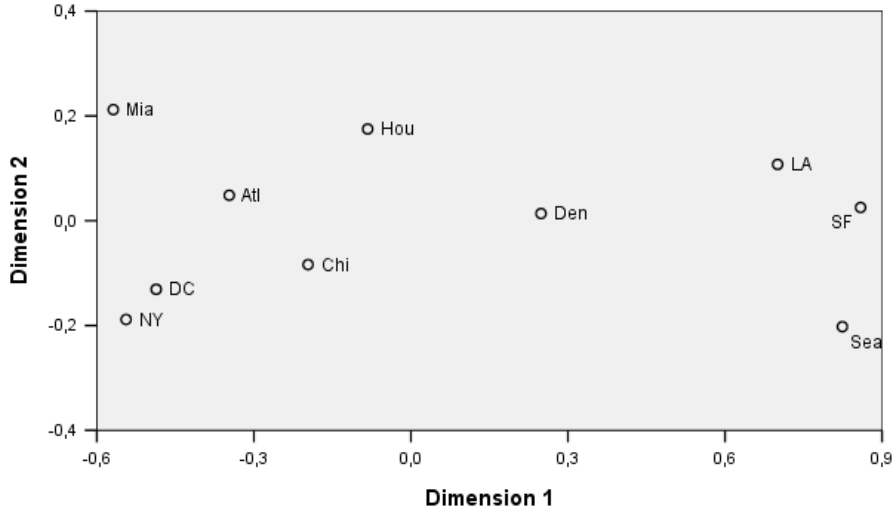
Abbildung 12-45
Korrespondenztabelle für Ähnlichkeiten

	Spalte										
Zeile	Atlanta	Chicago	Denver	Houston	Los Angeles	Miami	New York	San Francisco	Seattle	Washington DC	Aktiver Rand
Atl	2734	2147	1522	2033	798	2130	1986	595	552	2191	16688
Chi	2147	2734	1814	1794	989	1546	2021	876	997	2137	17055
Den	1522	1814	2734	1855	1903	1008	1103	1785	1713	1240	16677
Hou	2033	1794	1855	2734	1360	1766	1314	1089	843	1514	16302
LA	798	989	1903	1360	2734	395	283	2387	1775	434	13058
Mia	2130	1546	1008	1766	395	2734	1642	140	0	1811	13172
NY	1986	2021	1103	1314	283	1642	2734	163	326	2529	14101
SF	595	876	1785	1089	2387	140	163	2734	2056	292	12117
See	552	997	1713	843	1775	0	326	2056	2734	405	11401
DC	2191	2137	1240	1514	434	1811	2529	292	405	2734	15287
Aktiver Rand	16688	17055	16677	16302	13058	13172	14101	12117	11401	15287	145858

Zeilen- und Spaltenwerte

Durch die Verwendung von Flugmeilen anstelle von Straßenmeilen wirken sich die Gegebenheiten am Boden der USA nicht auf die Distanzen aus. Alle Ähnlichkeiten sollten daher in nur zwei Dimensionen dargestellt werden können. Zentrieren Sie sowohl die Zeilen als auch die Spalten und verwenden Sie die Prinzipal-Normalisierung. Aufgrund der Symmetrie der Korrespondenztabelle und der Prinzipal-Normalisierung sind die Zeilen- und Spaltenwerte identisch und die Gesamtträgheit befindet sich sowohl in den Zeilen als auch in den Spalten. Aus diesem Grund können Sie wahlweise die Zeilen- oder Spaltenwerte untersuchen und dabei jeweils dieselben Ergebnisse erzielen.

Abbildung 12-46
Punkte für 10 Städte



Die Positionen der Städte besitzen große Ähnlichkeit mit der tatsächlichen geografischen Lage, gedreht um den Ursprung. Weiter südlich liegende Städte zeigen größere Werte in der zweiten Dimension, weiter westlich liegende Städte dagegen größere Werte in der ersten Dimension.

Empfohlene Literatur

In folgenden Texten finden Sie weitere Informationen zur Korrespondenzanalyse:

Fisher, R. A. 1938. *Statistical methods for research workers*. Edinburgh: Oliver and Boyd.

Fisher, R. A. 1940. The precision of discriminant functions. *Annals of Eugenics*, 10, 422–429.

Gilula, Z., als auch S. J. Haberman. 1988. The analysis of multivariate contingency tables by restricted canonical and restricted association models. *Journal of the American Statistical Association*, 83, 760–771.

Kapitel
13

Mehrfachkorrespondenzanalyse

Der Zweck der Mehrfachkorrespondenzanalyse, die auch als Homogenitätsanalyse bekannt ist, besteht darin, Quantifizierungen zu finden, die in dem Sinne optimal sind, daß die Kategorien so weit wie möglich voneinander getrennt sind. Dies bringt mit sich, daß Objekte innerhalb derselben Kategorie eng beieinander und Objekte in unterschiedlichen Kategorien möglichst weit entfernt voneinander dargestellt werden. Der Ausdruck **Homogenität** bezieht sich auch auf die Tatsache, daß die Analyse am erfolgreichsten ausfällt, wenn die Variablen homogen sind, d. h. wenn sie die Objekte in Klumpen mit der gleichen Kategorie oder ähnlichen Kategorien aufteilen.

Beispiel: Eigenschaften von Eisenwaren

Um festzustellen, wie die Mehrfachkorrespondenzanalyse funktioniert, verwenden Sie Daten aus Hartigan (Hartigan, 1975), die Sie in der Datei *screws.sav* finden. Für weitere Informationen siehe Beispieldateien in Anhang A auf S. 326. Dieses Daten-Set enthält Informationen zu den Eigenschaften von Schrauben, Bolzen, Muttern und Reißnägeln. Die folgende Tabelle zeigt die Variablen mit den zugehörigen Variablenlabels und den Wertelabels, die den Kategorien jeder Variablen im Daten-Set für Eisenwaren von Hartigan zugeordnet sind.

Tabelle 13-1
Daten-Set für Eisenwaren von Hartigan

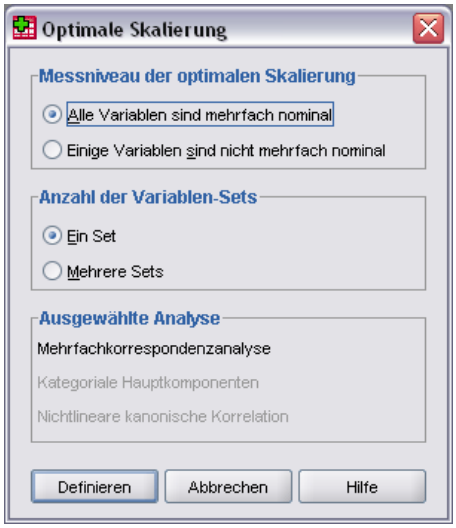
Variablenname	Variablenlabel	Wertelabels
gewinde	*Gewinde*	*Ja*, *Nein*
kopf	*Head form (Kopfform)*	*flach*, *konisch*, *rund*, *gewölbt*, *zylinderförmig*
kerbe	*Einkerbung des Kopfes*	*Schlitz*, *keine*, *Kreuz*
form	*Grundform*	*flach*, *scharf*
länge	*Länge in cm*	*0,5 cm*, *1 cm*, *1,5 cm*, *2 cm*, *2,5 cm*
messing	*Messing*	*Nein*, *Ja*
object	*Objekt*	*Reißnagel, Nagel1, Nagel2, Nagel3, Nagel4, Nagel5, Nagel6, Nagel7, Nagel8, Schraube1, Schraube2, Schraube3, Schraube4, Schraube5, Bolzen1, Bolzen2, Bolzen3, Bolzen4, Bolzen5, Bolzen6, Nagel1, nagel2, Nagelb, Schraubeb*

Durchführung der Analyse

▶ Zum Erstellen einer Mehrfachkorrespondenzanalyse wählen Sie die folgenden Befehle aus den Menüs aus:

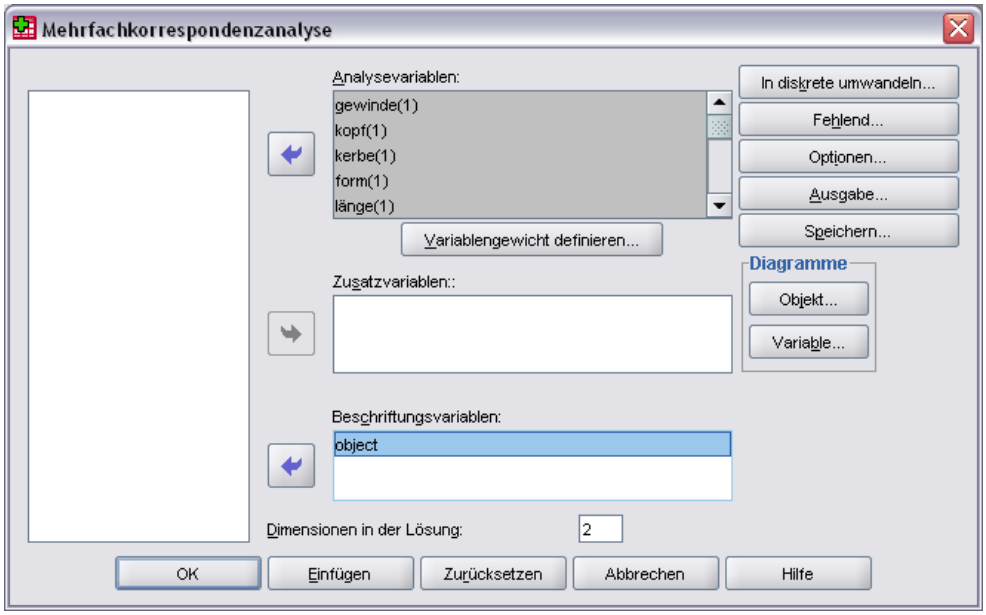
Analysieren
  Dimensionsreduktion
    Optimale Skalierung...

Abbildung 13-1
Dialogfeld "Optimale Skalierung"



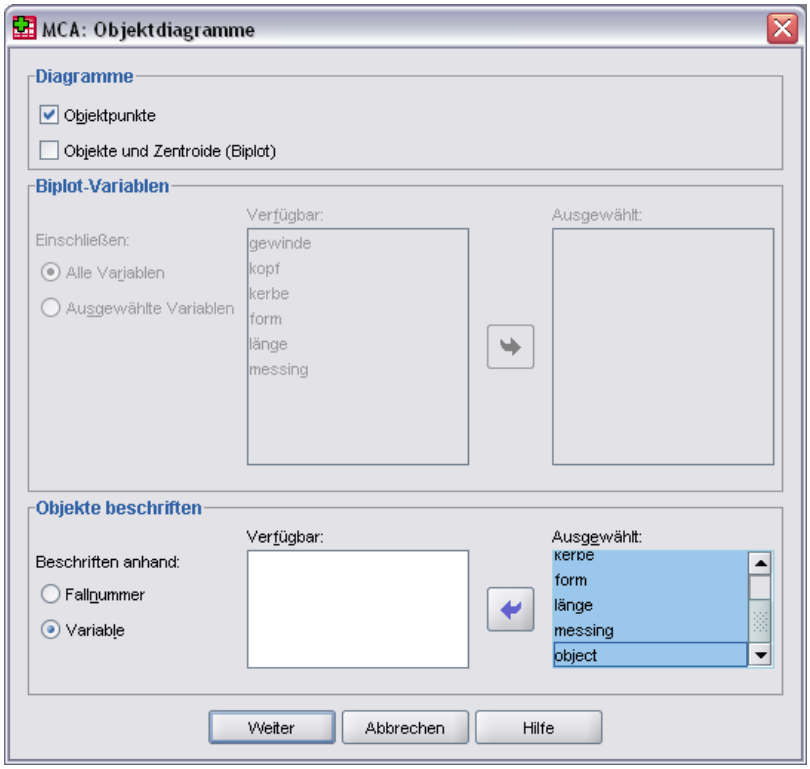
▶ Vergewissern Sie sich, daß Alle Variablen sind mehrfach nominal und Ein Set ausgewählt sind, und klicken Sie auf Definieren.

Abbildung 13-2
Dialogfeld "Mehrfachkorrespondenzanalyse"



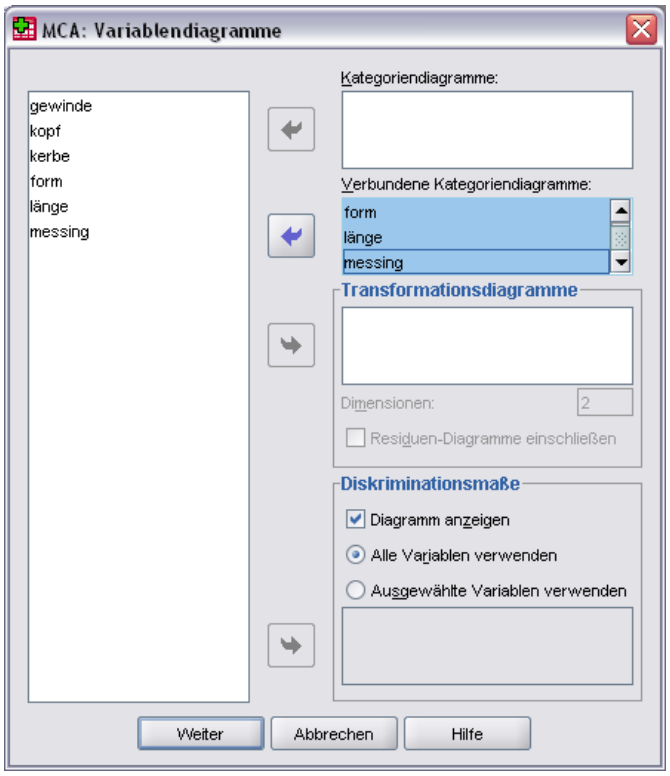
- Wählen Sie *Gewinde* bis *Länge in cm* als Analysevariablen aus.
- Wählen Sie *object* als Beschriftungsvariable.
- Klicken Sie im Gruppenfeld "Diagramme" auf Objekt.

Abbildung 13-3
Dialogfeld "Objektdiagramme"



- Wählen Sie aus, daß Objekte nach Variable beschriftet werden sollen.
- Wählen Sie *gewinde* bis *object* als Beschriftungsvariablen aus.
- Klicken Sie auf Weiter und dann im Gruppenfeld "Diagramme" des Dialogfelds "Mehrfachkorrespondenzanalyse" auf Variable.

Abbildung 13-4
Dialogfeld "Variablendiagramme"



- Wählen Sie aus, daß ein verbundenes Kategoriendiagramm für *gewinde* bis *länge* erstellt werden soll.
- Klicken Sie auf Weiter.
- Klicken Sie im Dialogfeld "Mehrfachkorrespondenzanalyse" auf OK.

Modellzusammenfassung

Mit der Homogenitätsanalyse kann eine Lösung für mehrere Dimensionen berechnet werden. Die maximale Anzahl an Dimensionen ist entweder die Anzahl der Kategorien abzüglich der Anzahl der Variablen ohne fehlende Daten oder die Anzahl der Beobachtungen minus 1, je nachdem, welche Zahl kleiner ist. Die maximale Anzahl an Dimensionen sollte jedoch nur selten verwendet werden. Eine kleinere Anzahl an Dimensionen läßt sich leichter interpretieren. Außerdem wird nach einer gewissen Anzahl an Dimensionen der Grad an Zusammenhang, der zusätzlich erklärt wird, vernachlässigbar. Eine ein-, zwei oder dreidimensionale Lösung bei der Homogenitätsanalyse ist sehr häufig.

Abbildung 13-5
Modellzusammenfassung

Dimension	Cronbachs Alpha	Varianz berücksichtigt		
		Gesamt (Eigenwert)	Trägheit	% der Varianz
1	,878	3,727	,621	62,123
2	,657	2,209	,368	36,809
Gesamt		5,936	,989	
Mittelwert	,796[a]	2,968	,495	49,466

a. Der Mittelwert von Cronbachs Alpha basiert auf dem Mittelwert der Eigenwerte.

Fast die gesamte Varianz der Daten wird durch die Lösung erklärt, 62,1 % durch die erste Dimension und 36,8 % durch die zweite.

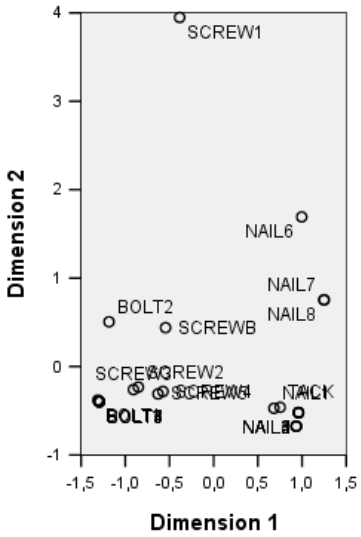
Zusammen bieten die beiden Dimensionen eine Interpretation hinsichtlich der Distanzen. Wenn Variablen gute Unterscheidungen bieten, befinden sich die Objekte nahe bei den Kategorien, zu denen sie gehören. Idealerweise befinden sich Objekte in derselben Kategorie nahe beieinander (sie sollten also ähnliche Bewertungen aufweisen) und Kategorien mit unterschiedlichen Variablen befinden sich nahe beieinander, wenn Sie zu denselben Objekten gehören (d. h. zwei Objekte, die ähnliche Werte für eine bestimmte Variable aufweisen, sollten auch ähnliche Bewertungen für die anderen Variablen in der Lösung aufweisen).

Objektwerte

Nach der Untersuchung der Modellzusammenfassung, sollten Sie sich den Objektwerten zuwenden. Sie können eine oder mehrere Variablen für die Beschriftung des Objektwertediagramms angeben. Jede Beschriftungsvariable erzeugt ein separates Diagramm, das mit den Werten dieser Variablen beschriftet ist. Wir betrachten das Diagramm der Objektwerte, das anhand des Variablenobjekts beschriftet ist. Dies ist nur eine Fallidentifizierungsvariable, die nicht bei Berechnungen verwendet wurde.

Die Distanz von einem Objekt bis zum Ursprung zeigt die Variation gegenüber dem "durchschnittlichen" Antwortmuster an. Dieses durchschnittliche Antwortmuster entspricht der häufigsten Kategorie für jede Variable. Objekte, bei denen viele Merkmale den häufigsten Kategorien entsprechen, liegen in der Nähe des Ursprungs. Objekte mit einzigartigen Merkmalen befinden sich hingegen weit vom Ursprung entfernt.

Abbildung 13-6
Objektwertediagramm, mit Objekt beschriftet



Bei der Untersuchung des Diagramms stellen wir fest, dass die erste Dimension (horizontale Achse) eine Unterscheidung zwischen Schrauben und Bolzen (weisen ein Gewinde auf) einerseits und Nägeln und Reißnägeln (kein Gewinde) andererseits bietet. Dies ist im Diagramm ganz einfach zu erkennen, da sich die Schrauben und Bolzen an einem Ende der horizontalen Achse befinden und die Nägel und Reißzwecken am anderen. In geringerer Schärfe trennt die erste Dimension auch die Bolzen (mit flacher Grundform) von allen anderen (mit spitzer Grundform).

Die zweite Dimension (vertikale Achse) scheint *SCHRAUBE1* und *NAGEL6* von allen anderen Objekten zu trennen. Die Gemeinsamkeit zwischen *SCHRAUBE1* und *NAGEL6* besteht in ihren Werten bei der Variablen "länge": Sie sind die längsten Objekte in der Datenmenge. Außerdem liegt *SCHRAUBE1* wesentlich weiter vom Ursprung entfernt als die anderen Objekte, was darauf hindeutet, daß insgesamt viele der Merkmale dieses Objekts bei den anderen Objekten nicht vorkommen.

Das Objektwertediagramm ist besonders hilfreich für die Ermittlung von Ausreißern. *SCHRAUBE1* könnte als Ausreißer betrachtet werden. Später werden wir untersuchen, was passiert, wenn wir dieses Objekt ausschließen.

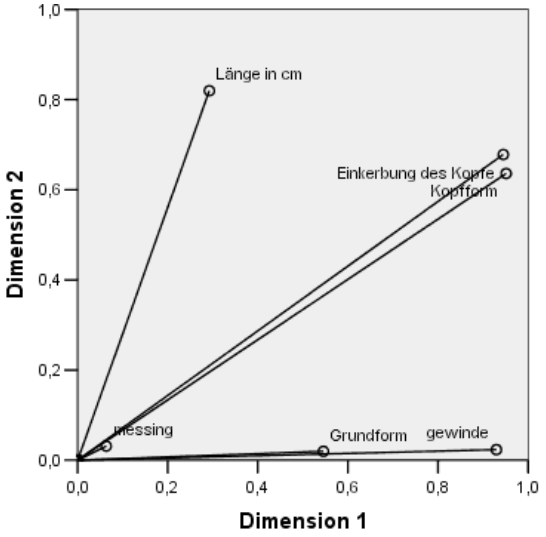
Diskriminationsmaße

Bevor wir den Rest der Objektwertediagramme untersuchen, wollen wir überprüfen, ob die Diskriminationsmaße mit unseren bisherigen Aussagen übereinstimmen. Bei jeder Variablen wird für jede Dimension ein Diskriminationsmaß, das als quadrierte Komponentenladung betrachtet werden kann, berechnet. Dieses Maß ist auch die Varianz der quantifizierten Variablen in dieser Dimension. Der Höchstwert für dieses Maß ist 1. Dieser Wert wird erreicht, wenn die Objektwerte in sich gegenseitig ausschließenden Gruppen liegen und alle Objektwerte innerhalb einer Kategorie gleich sind. (*Anmerkung*: Wenn fehlende Daten vorliegen, kann dieses Maß größer als 1 sein.) Große Diskriminationsmaße entsprechen einer großen Streubreite über die

Kategorien der Variablen und zeigen schließlich einen hohen Grad an Diskrimination zwischen den Kategorien einer Variablen entlang dieser Dimension an.

Der Durchschnitt der Diskriminationsmaße für eine Dimension ist gleich dem Prozentsatz der Varianz, die für diese Dimension erklärt werden kann. Folglich sind die Dimensionen nach durchschnittlicher Diskrimination geordnet. Die erste Dimension weist die größte durchschnittliche Diskrimination auf, die zweite Dimension weist die zweitgrößte durchschnittliche Diskrimination auf usw. Auf diese Weise werden alle Dimensionen in der Lösung geordnet.

Abbildung 13-7
Diagramm der Diskriminationsmaße



Wie beim Objektwertediagramm angezeigt, zeigt das Diagramm der Diskriminationsmaße an, daß die erste Dimension mit den Variablen *Gewinde* und *Grundform* in Zusammenhang steht. Diese Variablen weisen große Diskriminationsmaße in der ersten Dimension und kleine Diskriminationsmaße in der zweiten Dimension auf. Daher weisen für diese beiden Variablen die Kategorien nur in der ersten Dimension eine große Streubreite auf. *Länge in cm* weist einen großen Wert in der zweiten Dimension auf, jedoch einen kleinen Wert in der ersten Dimension. Als Ergebnis liegt *Länge* am nächsten an der zweiten Dimension, was mit der Beobachtung aus dem Objektwertediagramm übereinstimmt, daß die zweite Dimension die längsten Objekte vom Rest zu trennen scheint. *Einkerbung des Kopfes* und *Kopfform* weisen in beiden Dimensionen relative große Werte auf, was eine Diskrimination sowohl in der ersten als auch in der zweiten Dimension anzeigt. Die Variable *Messing*, die sehr nah am Ursprung liegt, bietet in den ersten beiden Dimensionen überhaupt keine Unterscheidung. Dies ergibt einen Sinn, da alle Objekte aus Messing oder auch nicht aus Messing bestehen können.

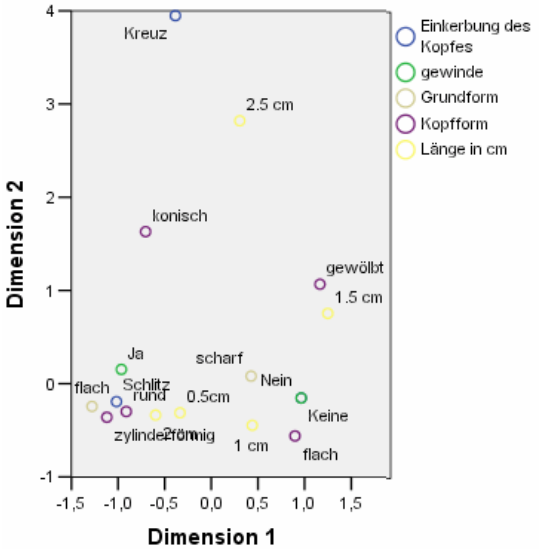
Kategorienquantifikationen

Erinnern wir uns, daß ein Diskriminationsmaß die Varianz der quantifizierten Variablen entlang einer bestimmten Dimension ist. Das Diagramm der Diskriminationsmaße enthält diese Varianzen, wodurch angezeigt wird, welche Variablen entlang welcher Dimension eine Unterscheidung

bieten. Dieselbe Varianz könnte jedoch vorliegen, wenn alle Kategorien mäßig weit voneinander entfernt liegen, oder wenn die meisten Kategorien eng zusammen liegen, einige Kategorien jedoch von dieser Gruppe abweichen. Das Diskriminationsdiagramm bietet keine Unterscheidung zwischen diesen beiden Zuständen.

Diagramme zur Kategorienquantifikation bieten eine alternative Methode zur Anzeige der Diskrimination von Variablen, durch die Kategoriebeziehungen identifiziert werden können. In diesem Diagramm werden die Koordinaten der einzelnen Kategorien in jeder Dimension angezeigt. So können Sie bestimmen, welche Kategorien für welche Variable ähnlich sind.

Abbildung 13-8
Kategorienquantifikationen



Länge in cm weist fünf Kategorien auf, von denen drei im oberen Bereich des Diagramms gruppiert sind. Die beiden anderen Kategorien befinden sich in der unteren Hälfte des Diagramms, wobei die Kategorie *2,5 cm* sehr weit von der Gruppe entfernt liegt. Die große Diskrimination für die Länge entlang Dimension 2 ergibt sich aus der Tatsache, daß sich diese eine Kategorie sehr stark von den anderen Längenkategorien unterscheidet. Ähnlich liegt bei *Kopfform* die Kategorie *KREUZ* sehr weit von den anderen Kategorien entfernt und bietet ein großes Diskriminationsmaß entlang der zweiten Dimension. Diese Muster können nicht in einem Diagramm der Diskriminationsmaße dargestellt werden.

Die Verteilung der Kategorienquantifikationen für eine Variable berücksichtigt die Varianz und zeigt so an, wie gut diese Variable in den einzelnen Dimensionen diskriminiert ist. In Dimension 1 liegen die Kategorien für *Gewinde* weit auseinander. Entlang Dimension 2 jedoch liegen die Kategorien für diese Variable sehr eng beieinander. *Gewinde* weist also in Dimension 1 eine bessere Diskrimination auf als in Dimension 2. Die Kategorien für *Kopfform* sind in beiden Dimensionen weit verstreut, was nahelegt, dass diese Variable in beiden Dimensionen eine gute Diskrimination bietet.

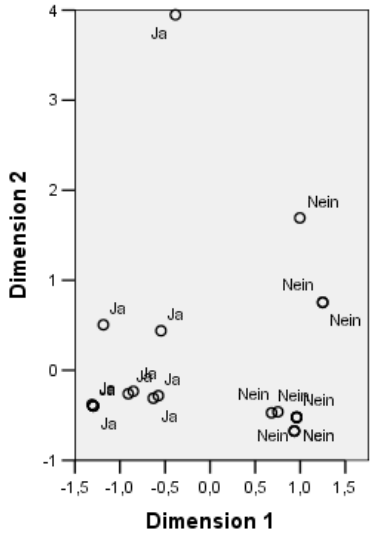
Zusätzlich zur Bestimmung der Dimensionen, entlang denen eine Variable Diskriminationen bietet, und der Art und Weise, in der sie diskriminiert, vergleicht das Diagramm der Kategoriequantifikationen auch die Variablendiskrimination. Eine Variable mit Kategorien, die weit auseinander liegen, bietet eine bessere Diskrimination als eine Variable mit Kategorien,

die eng beieinander liegen. So liegen entlang Dimension 1 die beiden Kategorien von *Messing* wesentlich enger beieinander als die beiden Kategorien von *Gewinde*, was bedeutet, daß *Gewinde* entlang dieser Dimension eine bessere Diskrimination aufweist als *Messing*. Entlang Dimension 2 sind die Abstände jedoch sehr ähnlich, was nahelegt, daß diese Variablen entlang dieser Dimension bis zum selben Grad diskriminieren. Das weiter oben erörterte Diskriminationsmaß identifiziert diese Beziehungen, indem Varianzen verwendet werden, um die Verteilung der Kategorien widerzuspiegeln.

Genauere Untersuchung der Objektwerte

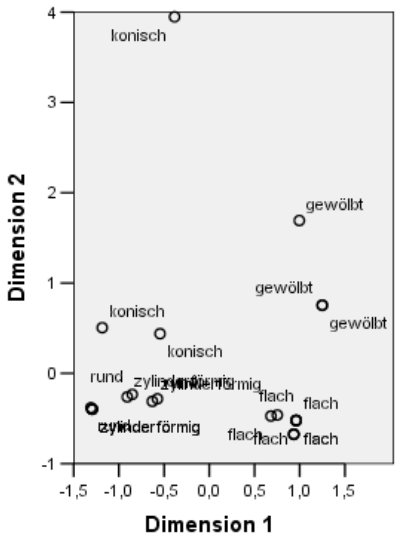
Tiefere Einblicke in die Daten können durch Untersuchung der nach den einzelnen Variablen beschrifteten Objektwertediagramme gewonnen werden. Idealerweise sollten ähnliche Objekte ausschließliche Gruppen bilden, und diese sollten weit voneinander entfernt liegen.

Abbildung 13-9
Objektwerte beschriftet mit "Gewinde"



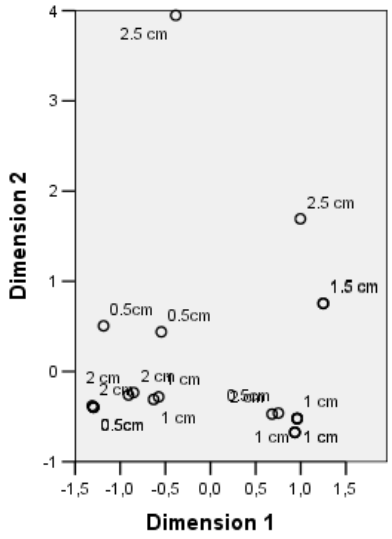
Das mit *Gewinde* beschriftete Diagramm zeigt, daß die erste Dimension *Ja*und *Nein* perfekt trennt. Alle Objekte mit Gewinde weisen negative Objektwerte auf, während alle Objekte ohne Gewinde positive Werte aufweisen. Die beiden Kategorien bilden zwar keine kompakten Gruppen, die perfekte Differenzierung zwischen den Kategorien wird jedoch allgemein als gutes Ergebnis betrachtet.

Abbildung 13-10
Objektwerte beschriftet mit "Kopfform"



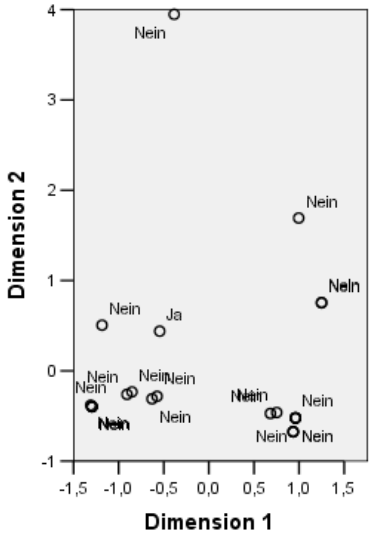
Das mit *Kopfform* beschriftete Diagramm zeigt, daß diese Variable in beiden Dimensionen diskriminiert. Die Objekte mit dem Wert *FLACH* bilden eine Gruppe in der rechten unteren Ecke des Diagramms, während die Objekte mit dem Wert *GEWÖLBT* eine Gruppe in der rechten oberen Ecke bilden. Die Objekte mit dem Wert *KONISCH* liegen alle oben links. Diese Objekte sind jedoch weiter verteilt als die anderen Gruppen und sind daher nicht so homogen. Schließlich können die Objekte mit dem Wert *ZYLINDERFÖRMIG* nicht von den Objekten mit dem Wert *RUND* getrennt werden, da beide in der linken unteren Ecke des Diagramms liegen.

Abbildung 13-11
Objektwerte beschriftet mit "Länge in cm"



Das mit *Länge in cm* beschriftete Diagramm zeigt, daß diese Variable in der ersten Dimension keine Diskrimination aufweist. Ihre Kategorien weisen bei einer Projektion auf eine horizontale Linie keine Gruppierung auf. In der zweiten Dimension jedoch liegt bei *Länge in cm* durchaus eine Diskrimination vor. Die kürzeren Objekte entsprechen positiven Werten, und die längeren Objekte entsprechen großen negativen Werten.

Abbildung 13-12
Objektwerte beschriftet mit "Messing"



Das mit *Messing* beschriftete Diagramm zeigt, dass diese Variable Kategorien aufweist, die sich in der ersten und zweiten Dimension nicht gut trennen lassen. Die Objektwerte liegen über den gesamten Raum verteilt. Die Objekte aus Messing können nicht von den Objekten differenziert werden, die nicht aus Messing bestehen.

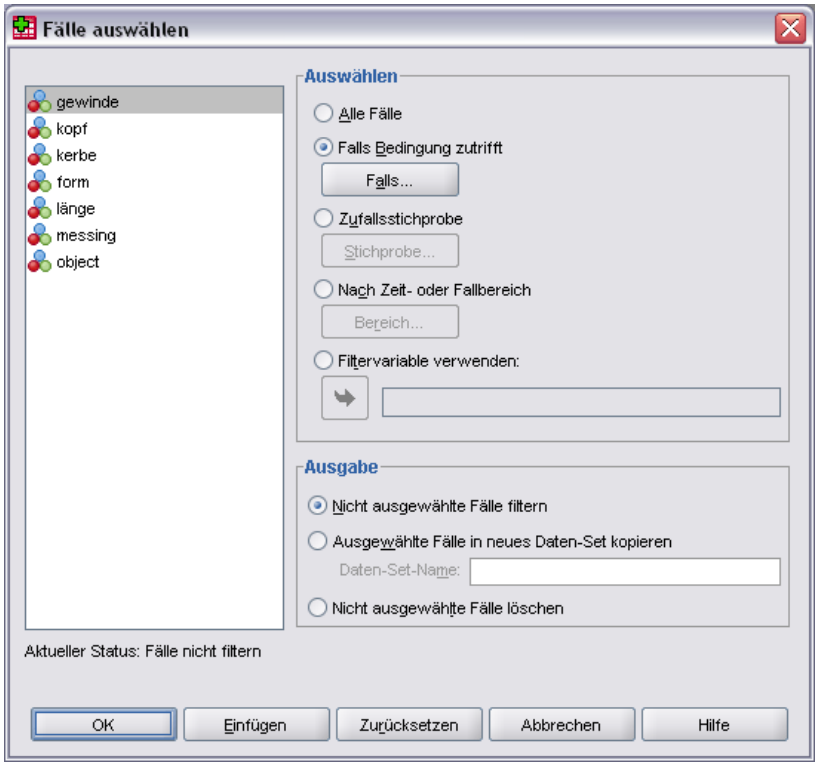
Ausschluß von Ausreißern

Bei der Homogenitätsanalyse sind Ausreißer Objekte, die zu viele einzigartige Merkmale aufweisen. Wie bereits erwähnt, könnte *SCHRAUBE1* als Ausreißer betrachtet werden.

Um dieses Objekt zu löschen und die Analyse erneut durchzuführen, wählen Sie folgende Optionen aus den Menüs aus:

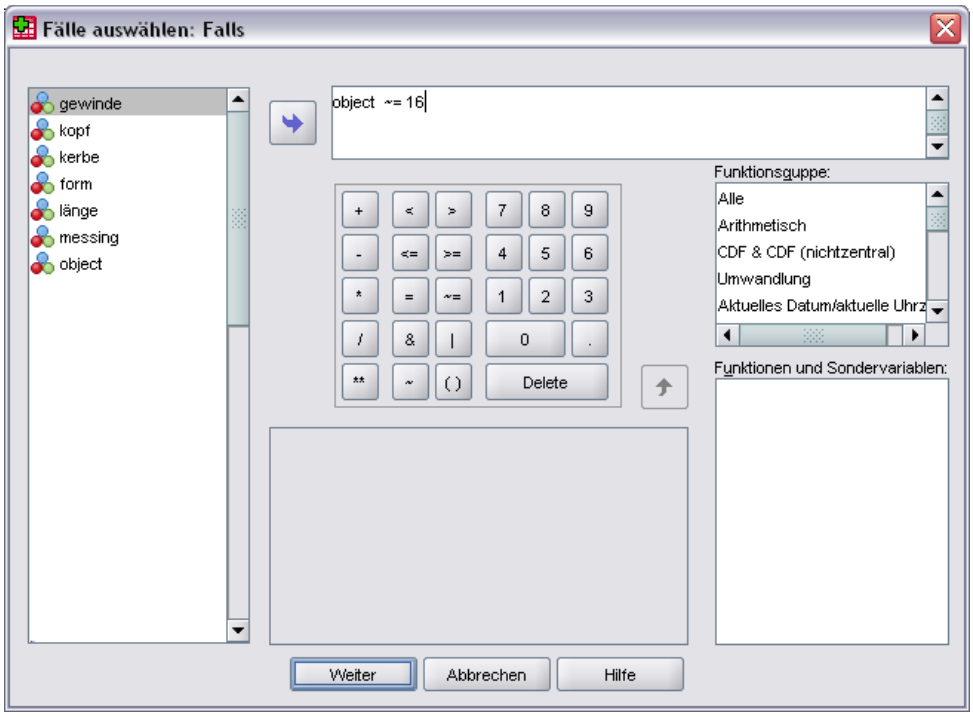
Daten
  Fälle auswählen...

Abbildung 13-13
Dialogfeld "Fälle auswählen"



- Aktivieren Sie das Optionsfeld Falls Bedingung zutrifft.
- Klicken Sie auf Falls.

Abbildung 13-14
Dialogfeld "Falls"



- Geben Sie object ~= 16 als Bedingung ein.
- Klicken Sie auf Weiter.
- Klicken Sie im Dialogfeld "Fälle auswählen" auf OK.
- Rufen Sie schließlich das Dialogfeld "Mehrfachkorrespondenzanalyse" auf, und klicken Sie auf OK.

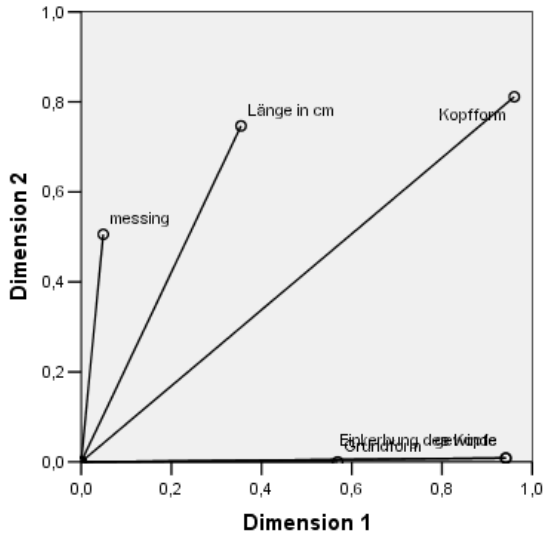
Abbildung 13-15
Modellzusammenfassung (nach der Entfernung des Ausreißers)

Dimension	Cronbachs Alpha	Varianz berücksichtigt		
		Gesamt (Eigenwert)	Trägheit	% der Varianz
1	,885	3,815	,636	63,591
2	,623	2,081	,347	34,676
Gesamt		5,896	,983	
Mittelwert	,793[a]	2,948	,491	49,133

a. Der Mittelwert von Cronbachs Alpha basiert auf dem Mittelwert der Eigenwerte.

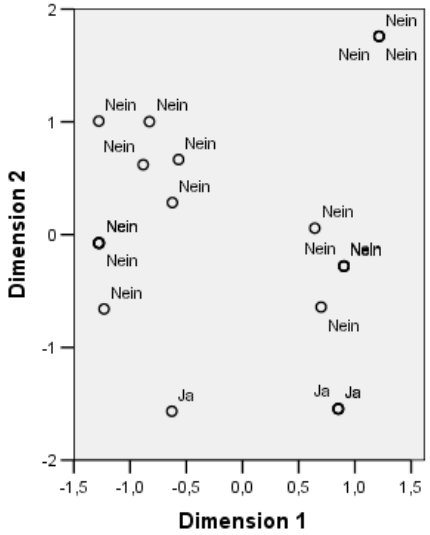
Die Eigenwerte sind leicht verschoben. Die erste Dimension erklärt nun einen etwas größeren Teil der Varianz.

Abbildung 13-16
Diskriminationsmaße



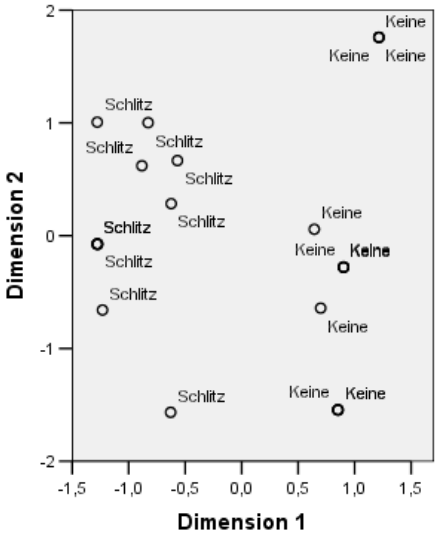
Wie im Diskriminationsdiagramm gezeigt, bietet *Einkerbung des Kopfes* keine Diskrimination in der zweiten Dimension mehr, während *Messing* nun nicht mehr wie zuvor gar keine Diskrimination aufweist, sondern zumindest in der zweiten Dimension eine Diskrimination bietet. Die Diskrimination für die übrigen Variablen bleibt weitgehend unverändert.

Abbildung 13-17
Objektwerte beschriftet mit "Messing" (nach der Entfernung des Ausreißers)



Das Objektwertediagramm mit der Beschriftung *Messing* zeigt, daß die vier Messingobjekte alle ziemlich weit unten im Diagramm erscheinen (drei Objekte nehmen exakt dieselbe Position ein), was eine hohe Diskrimination entlang der zweiten Dimension anzeigt. Wie auch bei *Gewinde* in der vorherigen Analyse der Fall, bilden die Objekte keine kompakten Gruppen, doch die Differenzierung der Objekte nach Kategorien ist perfekt.

Abbildung 13-18
Objektwerte beschriftet mit "Einkerbung des Kopfes" (nach der Entfernung des Ausreißers)



Das Objektwertediagramm mit der Beschriftung *Einkerbung des Kopfes* zeigt, daß die erste Dimension perfekt zwischen Objekten mit und ohne Einkerbung diskriminiert, wie auch in der vorherigen Analyse der Fall. Im Gegensatz zur vorherigen Analyse kann die zweite Dimension nun zwischen den beiden Kategorien unterscheiden.

Die Auslassung von *SCHRAUBE1*, dem einzigen Objekt mit einem Kreuzschlitz, hat erhebliche Auswirkungen auf die Interpretation der zweiten Dimension. Diese Dimension differenziert nun die Objekte auf der Grundlage von *Messing*, *Kopfform* und *Länge in cm*.

Empfohlene Literatur

In folgenden Texten finden Sie weitere Informationen zur Mehrfachkorrespondenzanalyse:

Benzécri, J. P. 1992. *Correspondence analysis handbook*. New York: Marcel Dekker.

Guttman, L. 1941. The quantification of a class of attributes: A theory and method of scale construction. In: *The Prediction of Personal Adjustment,* P. Horst (Hg.). New York: Social Science Research Council, 319–348.

Meulman, J. J. 1982. *Homogeneity analysis of incomplete data*. Leiden: DSWO Press.

Meulman, J. J. 1996. Fitting a distance model to homogeneous subsets of variables: Points of view analysis of categorical data. *Journal of Classification*, 13, 249–266.

Meulman, J. J., als auch W. J. Heiser. 1997. Graphical display of interaction in multiway contingency tables by use of homogeneity analysis. In: *Visual Display of Categorical Data,* M. Greenacre, als auch J. Blasius (Hgg.). New York: Academic Press, 277–296.

Nishisato, S. 1984. Forced classification: A simple application of a quantification method. *Psychometrika*, 49, 25–36.

Tenenhaus, M., als auch F. W. Young. 1985. An analysis and synthesis of multiple correspondence analysis, optimal scaling, dual scaling, homogeneity analysis, and other methods for quantifying categorical multivariate data. *Psychometrika*, 50, 91–119.

Van Rijckevorsel, J. 1987. *The application of fuzzy coding and horseshoes in multiple correspondence analysis*. Leiden: DSWO Press.

Kapitel
14

Multidimensionale Skalierung

Das Ziel der multidimensionalen Skalierung für eine gegebene Gruppe von Objekten besteht darin, eine Darstellung der Objekte in einem flachdimensionierten Raum zu finden. Diese Lösung wird mithilfe der **Ähnlichkeiten** zwischen den Objekten ermittelt. Bei der Prozedur werden die quadrierten Abweichungen zwischen den ursprünglichen (ggf. transformierten) Objektähnlichkeiten und den zugehörigen Euklidischen Distanzen im flachdimensionierten Raum minimiert.

Der flachdimensionierte Raum hat die Aufgabe, die Beziehungen zwischen den Objekten offenzulegen. Wenn Sie die Lösung so einschränken, dass nur eine lineare Kombination unabhängiger Variablen zulässig ist, können Sie die Dimensionen der Lösungen im Hinblick auf diese Variablen interpretieren. Im nachfolgenden Beispiel sehen Sie, wie 15 verschiedene Bezeichnungen des Verwandtschaftsgrades in drei Dimensionen dargestellt werden können und wie dieser Raum im Hinblick auf das Geschlecht, die Generation und den Trennungsgrad dieser Bezeichnungen interpretiert werden kann.

Beispiel: Untersuchung von Bezeichnungen des Verwandtschaftsgrades

Rosenberg und Kim (Rosenberg als auch Kim, 1975) haben 15 Bezeichnungen für den Verwandtschaftsgrad untersucht (Tante, Bruder, Cousin, Tochter, Vater, Enkelin, Großvater, Großmutter, Enkel, Mutter, Neffe, Nichte, Schwester, Sohn, Onkel). Die beiden Analytiker baten vier Gruppen von College-Studenten (zwei weibliche und zwei männliche Gruppen), diese Bezeichnungen auf der Grundlage der Ähnlichkeiten zu sortieren. Zwei Gruppen (eine weibliche und eine männliche Gruppe) wurden gebeten, die Bezeichnungen zweimal zu sortieren; die zweite Sortierung sollte dabei nach anderen Kriterien erfolgen als die erste. So wurden insgesamt sechs "Quellen" erzielt, die in der nachstehenden Tabelle dargestellt werden.

Tabelle 14-1
Quellstruktur der Verwandtschaftsdaten

Quelle	Geschlecht	Bedingung	Stichprobenumfang
1	Weiblich	Einfache Sortierung	85
2	Männlich	Einfache Sortierung	85
3	Weiblich	Erste Sortierung	80
4	Weiblich	Zweite Sortierung	80

Quelle	Geschlecht	Bedingung	Stichprobenumfang
5	Männlich	Erste Sortierung	80
6	Männlich	Zweite Sortierung	80

Jede Quelle entspricht einer Ähnlichkeitsmatrix mit 15×15 Elementen. Die Anzahl der Zellen ist dabei gleich der Anzahl der Personen in einer Quelle minus der Anzahl der Platzierungen der Objekte in diese Quelle. Dieses Daten-Set finden Sie in *kinship_dat.sav*. Für weitere Informationen siehe Beispieldateien in Anhang A auf S. 326.

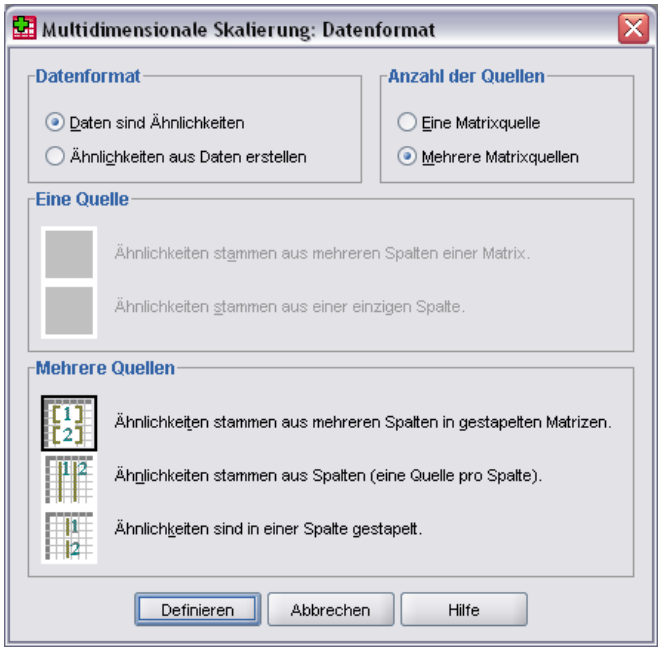
Festlegen der Anzahl der Dimensionen

Entscheiden Sie selbst, wie viele Dimensionen die Lösung erhalten soll. Ein Screeplot kann Sie bei dieser Entscheidung unterstützen.

▶ Zum Erstellen eines Screeplots wählen Sie die folgenden Befehle aus den Menüs aus:

Analysieren
  Skalieren
    Multidimensionale Skalierung (PROXSCAL)…

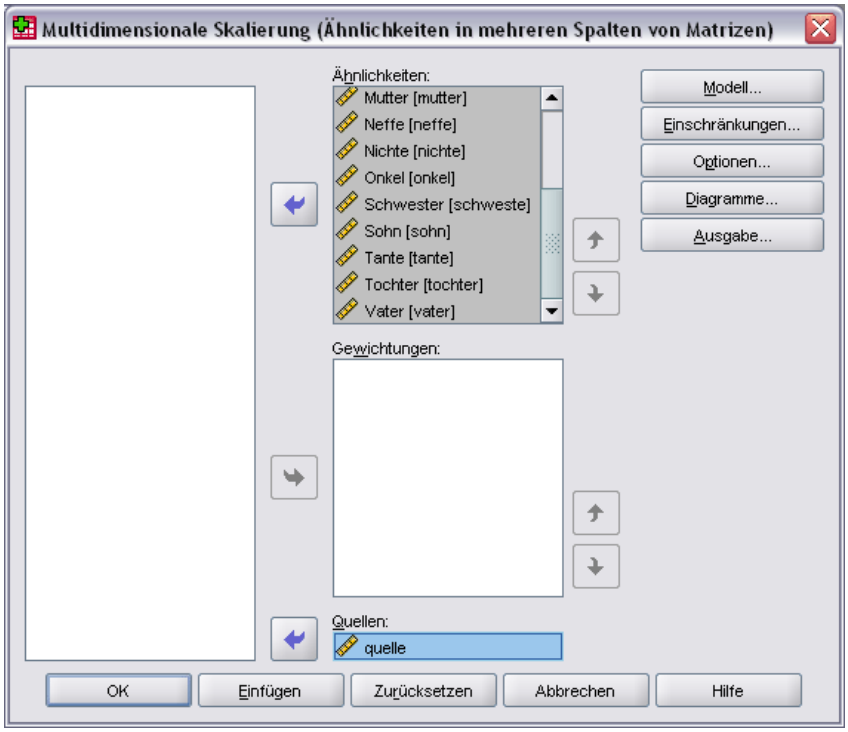
Abbildung 14-1
Dialogfeld "Datenformat"



▶ Wählen Sie in der Gruppe "Anzahl der Quellen" die Option Mehrere Matrixquellen.

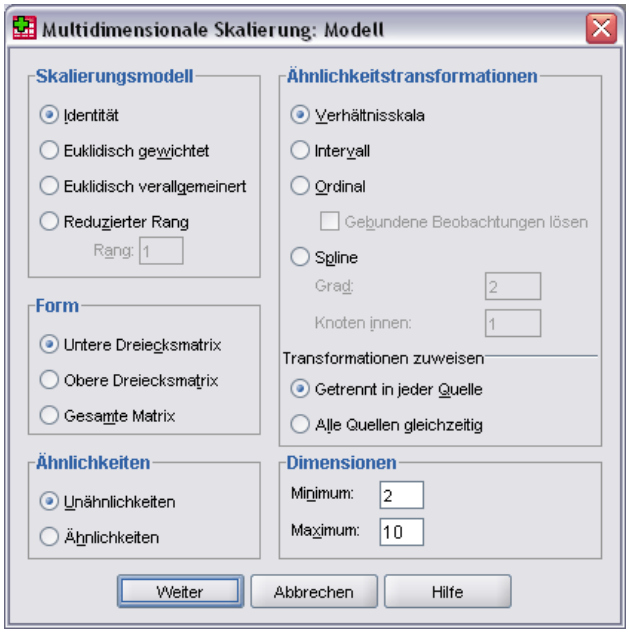
▶ Klicken Sie auf Definieren.

Abbildung 14-2
Dialogfeld "Multidimensionale Skalierung"



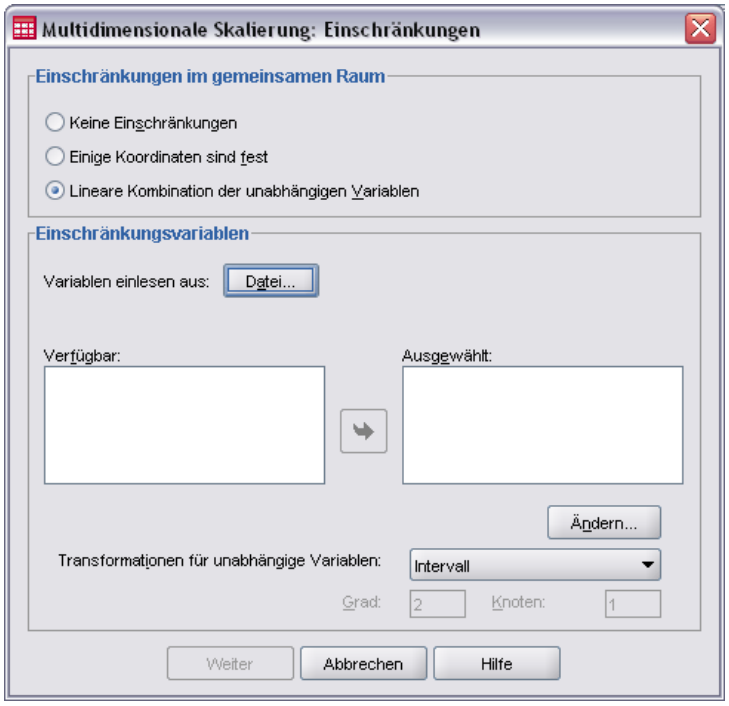
- Wählen Sie *Tante* bis *Onkel* als Ähnlichkeitsvariablen aus.
- Wählen Sie *quelle* als Variable für die Bezeichnung der Quelle aus.
- Klicken Sie auf Modell.

Abbildung 14-3
Dialogfeld "Modell"



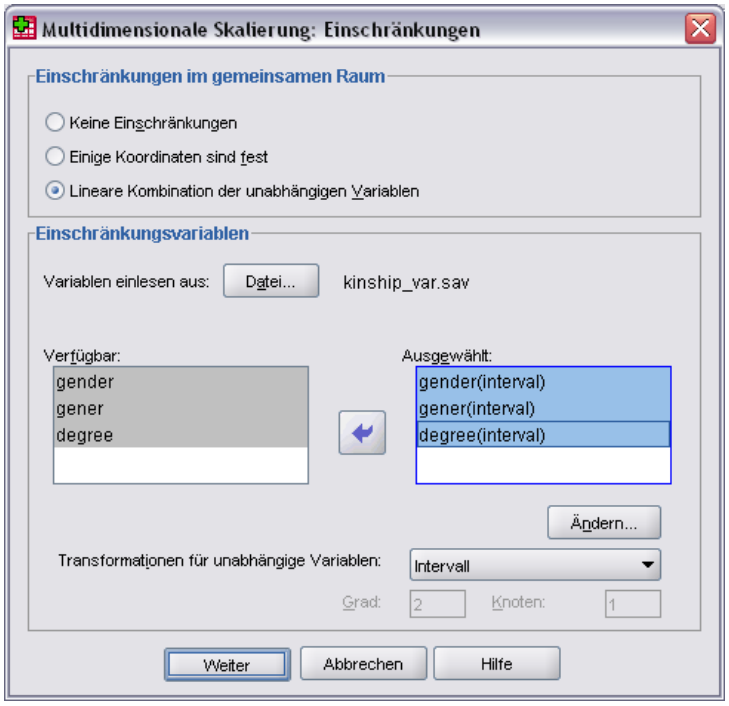
- Geben Sie 10 als maximal zulässige Anzahl der Dimensionen ein.
- Klicken Sie auf Weiter.
- Klicken Sie im Dialogfeld "Multidimensionale Skalierung" auf Einschränkungen.

Abbildung 14-4
Dialogfeld "Einschränkungen"



- Wählen Sie Lineare Kombination der unabhängigen Variablen.
- Klicken Sie auf Datei und wählen Sie die Quelle der unabhängigen Variablen aus.
- Wählen Sie die Datei *kinship_var.sav* aus.

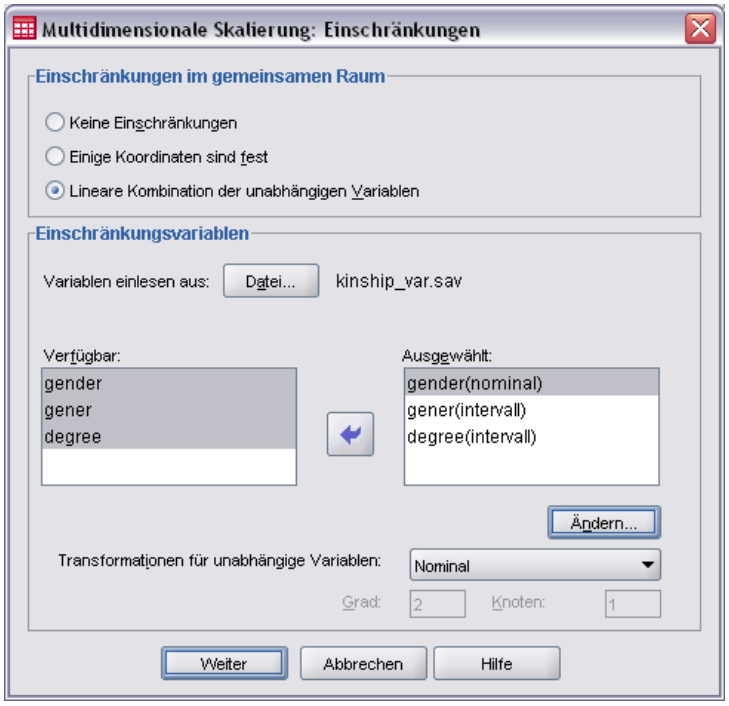
Abbildung 14-5
Dialogfeld "Einschränkungen"



▶ Wählen Sie *gender*, *gener* und *degree* als Einschränkungsvariablen aus.

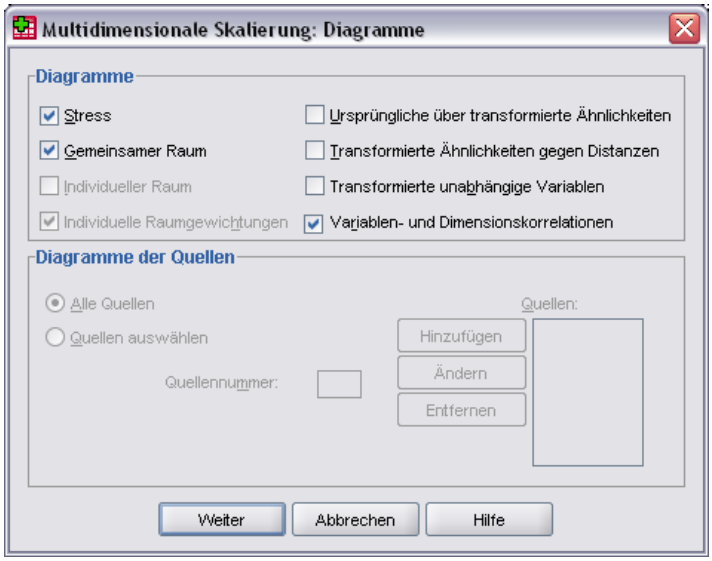
Beachten Sie, dass die Variable *gender* einen benutzerdefiniert fehlenden Wert aufweist — 9 = Fehlend (für "Cousin"). Im Rahmen der Prozedur wird dies als gültige Kategorie behandelt. Die lineare Transformation nach dem Standardverfahren ist daher eher ungeeignet. Verwenden Sie statt dessen eine nominale Transformation.

Abbildung 14-6
Dialogfeld "Einschränkungen"



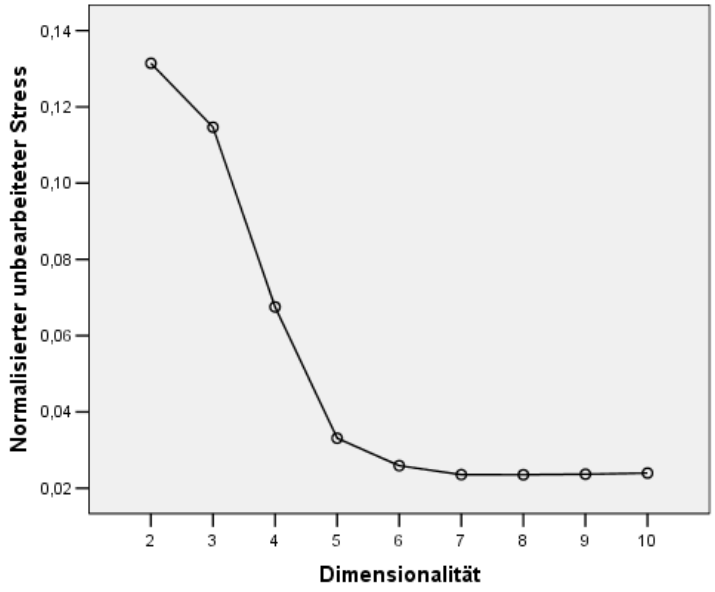
- Wählen Sie *gender* aus.
- Wählen Sie in der Dropdown-Liste für die Transformationen unabhängiger Variablen den Eintrag Nominal.
- Klicken Sie auf Ändern.
- Klicken Sie auf Weiter.
- Klicken Sie im Dialogfeld "Multidimensionale Skalierung" auf Diagramme.

Abbildung 14-7
Dialogfeld "Diagramme"



- Wählen Sie in der Gruppe "Diagramme" die Option Stress.
- Klicken Sie auf Weiter.
- Klicken Sie im Dialogfeld "Multidimensionale Skalierung" auf OK.

Abbildung 14-8
Screeplot



Die Prozedur beginnt mit einer zehndimensionalen Lösung und wird dann auf eine zweidimensionale Lösung reduziert. Der Screeplot zeigt den normalisierten Roh-Stress für die Lösung in jeder Dimension. Aus dem Diagramm geht hervor, dass ein Anstieg der Dimensionen

von 2 auf 3 sowie von 3 auf 4 eine große Verbesserung beim Stress nach sich zieht. Ab 4 werden nur noch geringfügige Verbesserungen erzielt. Sie entscheiden sich, die Daten mithilfe einer dreidimensionalen Lösung zu interpretieren, weil die Ergebnisse hier einfacher zu interpretieren sind.

Eine dreidimensionale Lösung

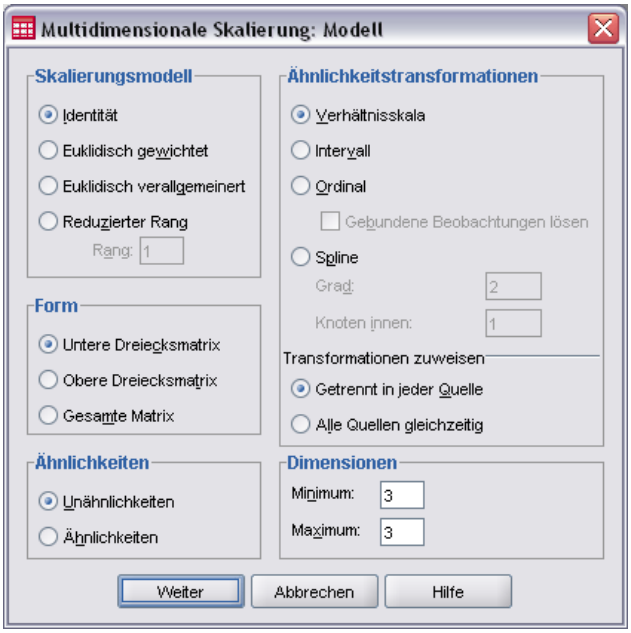
Die unabhängigen Variablen *gender*, *gener* (generation) und *degree* (der Trennung) wurden speziell für die Interpretation der Dimensionen in der Lösung erstellt. Die unabhängigen Variablen wurden wie folgt gebildet:

gender	1 = “male”, 2 = “female”, 9 = “missing” (für “Cousin”)
gener	Anzahl der Generationen, von Ihnen aus gerechnet, wenn diese Bezeichnung für Ihre Verwandtschaft verwendet wurde. Niedrigere Zahlen bedeuten dabei ältere Generationen. Großeltern erhalten den Wert –2, Enkel den Wert 2 und Geschwister den Wert 0.
degree	Die Anzahl der Trennungsgrade im Stammbaum. Ihre Eltern befinden sich 1 Knoten über Ihnen, Ihre Kinder entsprechend 1 Knoten unter Ihnen. Um zu Ihren Geschwistern zu gelangen, gehen Sie 1 Knoten nach oben zu Ihren Eltern und dann 1 Knoten abwärts zu Ihren Geschwistern. Hier liegen also 2 Trennungsgrade vor. Ihr Cousin ist 4 Trennungsgrade entfernt: —2 Knoten nach oben zu Ihren Großeltern, dann 2 Knoten nach unten (über Tante/Onkel) bis zum Cousin.

Die externen Variablen befinden sich in *kinship_var.sav*. Außerdem befindet sich eine Anfangskonfiguration aus einer früheren Analyse in *kinship_ini.sav*. Für weitere Informationen siehe Beispieldateien in Anhang A auf S. 326.

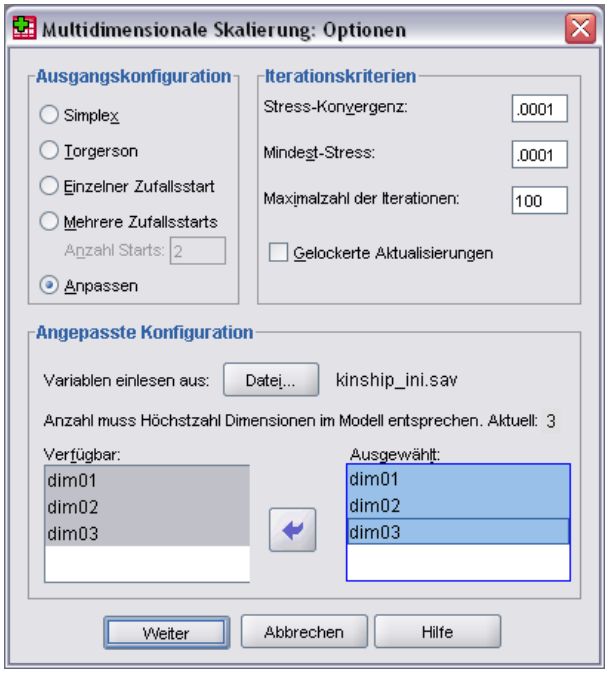
Durchführung der Analyse

Abbildung 14-9
Dialogfeld "Modell"



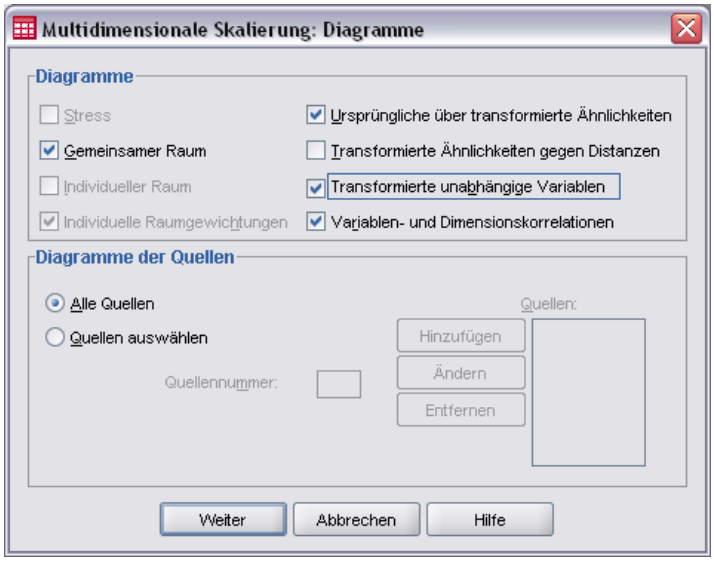
- Zum Erstellen einer dreidimensionalen Lösung öffnen Sie erneut das Dialogfeld "Multidimensionale Skalierung" und klicken Sie auf Modell.
- Geben Sie 3 als minimal und maximal zulässige Anzahl der Dimensionen ein.
- Klicken Sie auf Weiter.
- Klicken Sie im Dialogfeld "Multidimensionale Skalierung" auf Optionen.

Abbildung 14-10
Dialogfeld "Optionen"



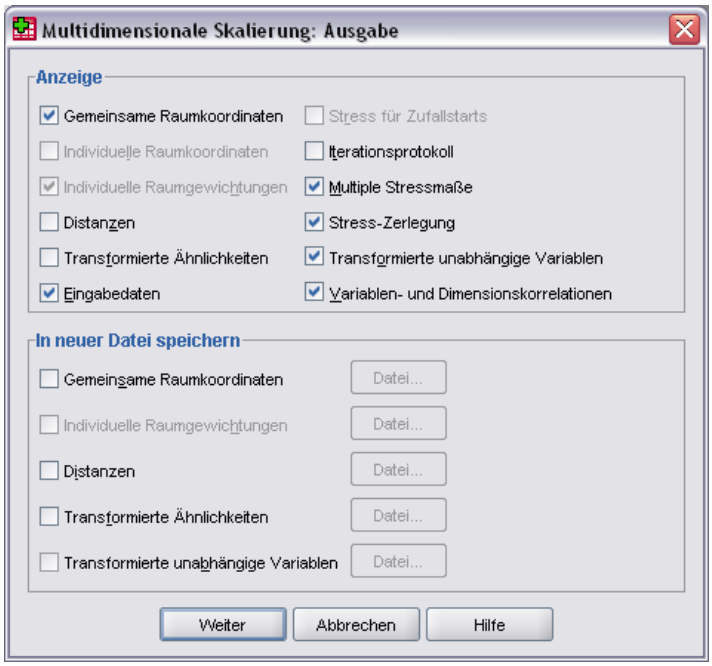
- Wählen Sie Anpassen als Anfangskonfiguration.
- Wählen Sie die Datei *kinship_ini.sav* als Datei, aus der die Variablen ausgelesen werden sollen.
- Wählen Sie *dim01*, *dim02* und *dim03* als Variablen aus.
- Klicken Sie auf Weiter.
- Klicken Sie im Dialogfeld "Multidimensionale Skalierung" auf Diagramme.

Abbildung 14-11
Dialogfeld "Diagramme"



- Wählen Sie Ursprüngliche über transformierte Ähnlichkeiten und Transformierte unabhängige Variablen.
- Klicken Sie auf Weiter.
- Klicken Sie im Dialogfeld "Multidimensionale Skalierung" auf Ausgabe.

Abbildung 14-12
Dialogfeld "Ausgabe"



- Wählen Sie Eingabedaten, Stress-Zerlegung und Variablen- und Dimensionskorrelationen.

- Klicken Sie auf **Weiter**.
- Klicken Sie im Dialogfeld "Multidimensionale Skalierung" auf **OK**.

Stressmaße

Die Stress- und Anpassungsmaße lassen erkennen, inwieweit sich die Distanzen in der Lösung den ursprünglichen Distanzen nähern.

Abbildung 14-13
Stress- und Anpassungsmaße

Normalisierter Roh-Stress	,06234
Stress-I	,24968[a]
Stress-II	,87849[a]
S-Stress	,14716[b]
Erklärte Streuung (D.A.F.)	,93766
Kongruenzkoeffizient nach Tucker	,96833

PROXSCAL minimiert den normalisierten Roh-Stress.

a. Faktor für optimale Skalierung = 1,066.

b. Faktor für optimale Skalierung = ,984.

Jedes der vier Stress-Statistikmaße misst die mangelnde Anpassung der Daten; die berücksichtigte Streuung und der Kongruenzkoeffizient nach Tucker messen dagegen die Anpassung. Niedrigere Stressmaße (bis zum Minimum 0) und höhere Anpassungsmaße (bis zum Maximum 1) weisen auf bessere Lösungen hin.

Abbildung 14-14
Zerlegung des normalisierten Roh-Stresses

		Quelle						
		SRC_1	SRC_2	SRC_3	SRC_4	SRC_5	SRC_6	Mittelwert
Objekt	Aunt	,0991	,0754	,0629	,0468	,0391	,0489	,0620
	Brother	,1351	,0974	,0496	,0813	,0613	,0597	,0807
	Cousin	,0325	,0336	,0480	,0290	,0327	,0463	,0370
	Daughter	,0700	,0370	,0516	,0229	,0326	,0207	,0391
	Father	,0751	,0482	,0521	,0225	,0272	,0298	,0425
	Granddaughter	,1410	,0736	,0801	,0707	,0790	,0366	,0802
	Grandfather	,1549	,1057	,0858	,0821	,0851	,0576	,0952
	Grandmother	,1550	,0979	,0858	,0844	,0816	,0627	,0946
	Grandson	,1374	,0772	,0793	,0719	,0791	,0382	,0805
	Mother	,0813	,0482	,0526	,0229	,0260	,0227	,0423
	Nephew	,0843	,0619	,0580	,0375	,0317	,0273	,0501
	Niece	,0850	,0577	,0503	,0353	,0337	,0260	,0480
	Sister	,1361	,0946	,0496	,0816	,0629	,0588	,0806
	Son	,0689	,0373	,0456	,0242	,0337	,0253	,0392
	Uncle	,0977	,0761	,0678	,0489	,0383	,0498	,0631
Mittelwert		,1035	,0681	,0613	,0508	,0496	,0407	,0623

Mit der Stress-Zerlegung erkennen Sie, welche Quellen und Objekte die größten Beiträge zum Gesamtstress der Lösung liefern. In diesem Fall kann der Großteil des Stresses zwischen den Quellen den Quellen 1 und 2 zugeordnet werden. Bei den Objekten läßt sich der meiste Stress dagegen mit *Bruder*, *Enkelin*, *Großvater*, *Großmutter*, *Enkel* und *Schwester* in Verbindung bringen.

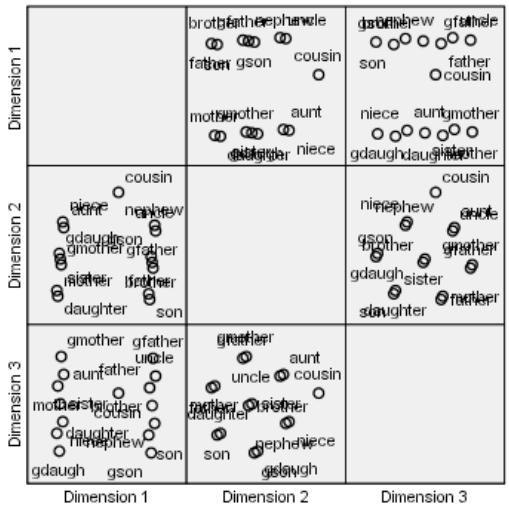
Die beiden Quellen, die den größten Teil des Stresses liefern, sind die beiden Gruppen, die die Bezeichnungen nur einmal sortiert hatten. Diese Information legt nahe, dass die Studenten beim Sortieren der Bezeichnungen mehrere Faktoren in Betracht gezogen hatten und dass sich die Studenten, die die Bezeichnungen zweimal sortierten, bei der ersten Sortierung auf einen Teil dieser Faktoren konzentrierten und dann die verbleibenden Faktoren bei der zweiten Sortierung anwendeten.

Die Objekte mit dem größten Anteil am Stress sind die Objekte mit einem Wert von 2 für *degree*, also Verwandte, die nicht zur Kernfamilie gehören (*Mutter*, *Vater*, *Tochter*, *Sohn*), aber dennoch näher verwandt sind als andere. Diese Mittelposition könnte leicht eine differentielle Sortierung dieser Terme verursachen.

Endgültige Koordinaten des gemeinsamen Raums

Im Diagramm mit dem gemeinsamen Raum werden die Beziehungen zwischen den Objekten anschaulich dargestellt.

Abbildung 14-15
Gemeinsame Raumkoordinaten



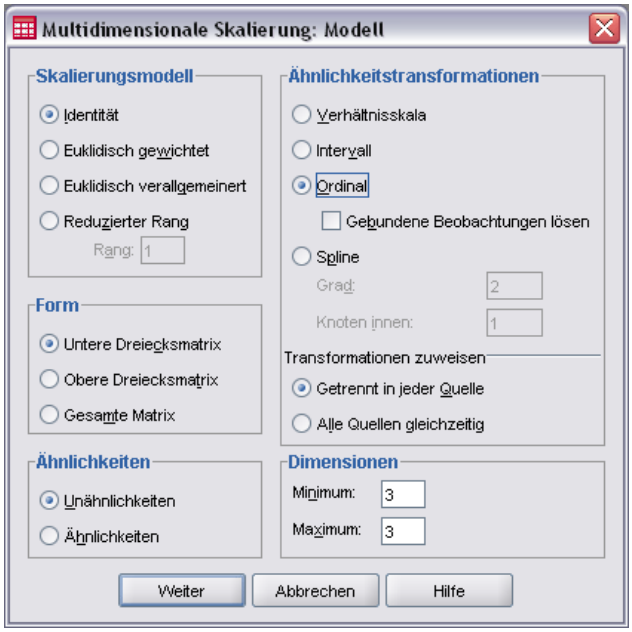
Betrachten Sie die endgültigen Koordinaten für die Objekte in den Dimensionen 1 und 3; dies ist das Diagramm im unteren linken Bereich der Streudiagramm-Matrix. Dieses Diagramm zeigt, dass Dimension 1 (auf der *x*-Achse) mit der Variablen *gender* korreliert, Dimension 3 (auf der *y*-Achse) dagegen mit der Variablen *gener*. Von links nach rechts wird ersichtlich, dass Dimension 1 die weiblichen und männlichen Bezeichnungen trennt; die im Grunde genommen geschlechtslose Bezeichnung *Cousin* befindet sich in der Mitte. Steigende Werte entlang der Achse von unten nach oben im Diagramm gehören zu Bezeichnungen älterer Familienmitglieder.

Betrachten Sie nun die endgültigen Koordinaten für die Objekte in den Dimensionen 2 und 3; dies ist das Diagramm im mittleren rechten Bereich der Streudiagramm-Matrix. Aus diesem Diagramm geht hervor, dass die zweite Dimension (auf der *y*-Achse) der Variablen *degree* entspricht; größere Werte an der Achse entsprechen dabei Bezeichnungen außerhalb der Kernfamilie.

Dreidimensionale Lösung mit nicht standardmäßigen Transformationen

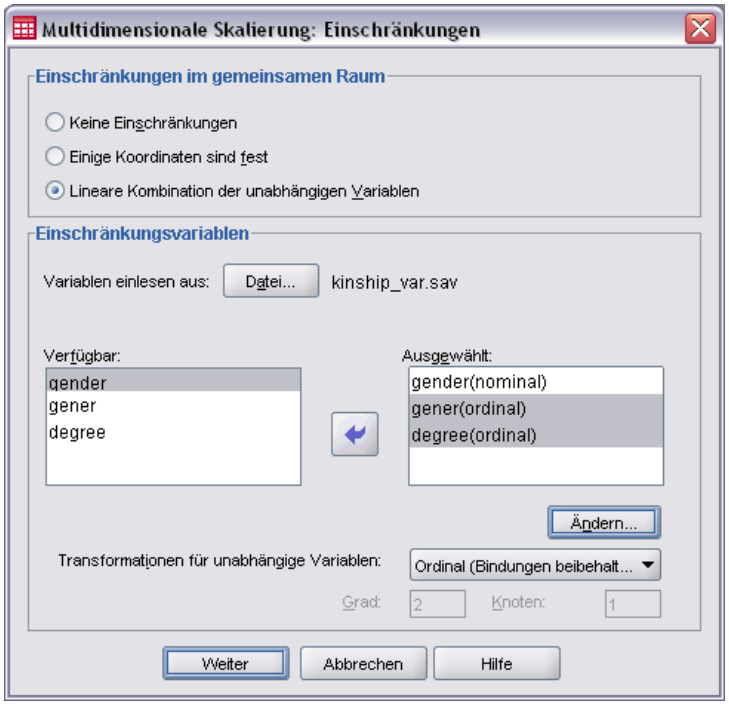
Bei der vorangegangenen Lösung wurde die Verhältnistransformation nach dem Standardverfahren für Ähnlichkeiten verwendet sowie Intervalltransformationen für die unabhängigen Variablen *gener* und *degree*. Die Ergebnisse sind schon recht gut; mit anderen Transformationen könnten Sie jedoch noch bessere Resultate erzielen. Die Ähnlichkeiten sowie die Variablen *gener* und *degree* besitzen beispielsweise jeweils eine natürliche Reihenfolge, könnten jedoch mit einer ordinalen Transformation noch besser modelliert werden als mit einer linearen Transformation.

Abbildung 14-16
Dialogfeld "Modell"



- Um die Analyse zu wiederholen und dabei die Ähnlichkeiten sowie die Variablen *gener* und *degree* auf einem ordinalen Messniveau neu zu skalieren (Bindungen werden beibehalten), öffnen Sie erneut das Dialogfeld "Multidimensionale Skalierung" und klicken Sie auf Modell.
- Wählen Sie Ordinal als Ähnlichkeitstransformation.
- Klicken Sie auf Weiter.
- Klicken Sie im Dialogfeld "Multidimensionale Skalierung" auf Einschränkungen.

Abbildung 14-17
Dialogfeld "Einschränkungen"



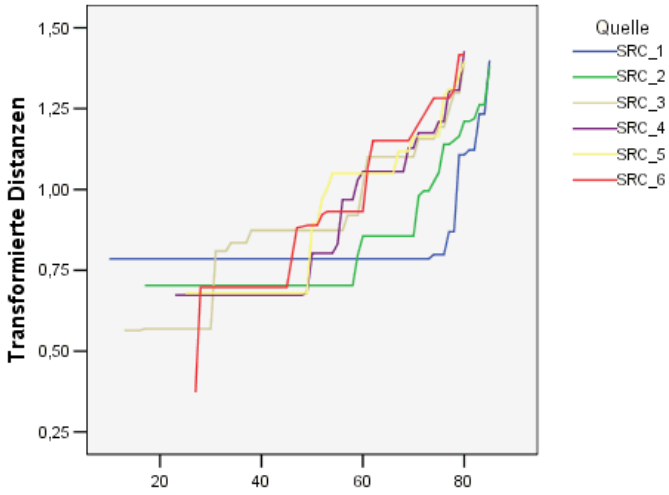
- Wählen Sie *gener* und *degree*.
- Wählen Sie in der Dropdown-Liste für die Transformationen unabhängiger Variablen den Eintrag Ordinal (Bindungen beibehalten).
- Klicken Sie auf Ändern.
- Klicken Sie auf Weiter.
- Klicken Sie im Dialogfeld "Multidimensionale Skalierung" auf OK.

Transformationsdiagramme

Mithilfe der Transformationsdiagramme können Sie rasch prüfen, ob die ursprünglichen Transformationen angemessen waren. Wenn die Diagramme ungefähr linear sind, ist die lineare Annahme geeignet. Prüfen Sie andernfalls die Stressmaße, ob eine Verbesserung bei der Anpassung vorliegt. Ermitteln Sie außerdem anhand des Diagramms mit dem gemeinsamen Raum, ob die Interpretation aussagekräftigere Ergebnisse liefert.

Die unabhängigen Variablen resultieren jeweils in einer ungefähr linearen Transformation; diese Variablen sollten daher ggf. als numerisch interpretiert werden. Die Ähnlichkeiten liefern allerdings keine lineare Transformation, sodass hier eine ordinale Transformation in Betracht gezogen werden sollte.

Abbildung 14-18
Transformierte Ähnlichkeiten



Stressmaße

Der Stress für die aktuelle Lösung stützt die Aussage, dass die Ähnlichkeiten auf einem ordinalen Messniveau skaliert werden sollten.

Abbildung 14-19
Stress- und Anpassungsmaße

Normalisierter Roh-Stress	,03137
Stress-I	,17712[a]
Stress-II	,61987[a]
S-Stress	,07953[b]
Erklärte Streuung (D.A.F.)	,96863
Kongruenzkoeffizient nach Tucker	,98419

PROXSCAL minimiert den normalisierten Roh-Stress.

a. Faktor für optimale Skalierung = 1,032.

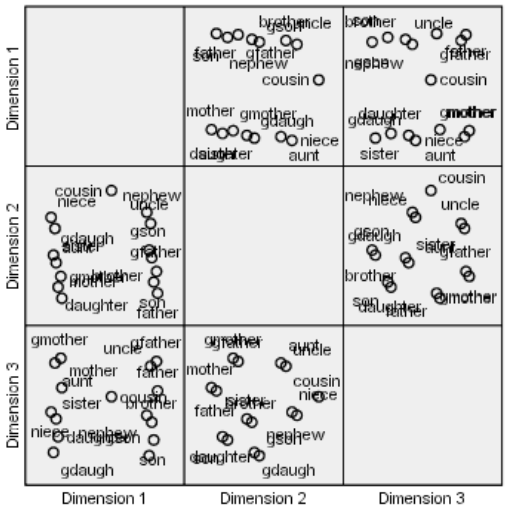
b. Faktor für optimale Skalierung = ,980.

Der normalisierte Roh-Stress für die vorangegangene Lösung liegt bei 0,06234. Durch die Skalierung der Variablen mit nicht standardmäßigen Transformationen wird der Wert für den Stress auf 0,03137 halbiert.

Endgültige Koordinaten des gemeinsamen Raums

Diagramme des gemeinsamen Raums ermöglichen im wesentlichen dieselbe Interpretation der Dimensionen wie die vorangegangene Lösung.

Abbildung 14-20
Koordinaten des gemeinsamen Raums



Diskussion

Die Ähnlichkeiten sollten am besten als ordinale Variablen behandelt werden, da eine deutliche Verbesserung bei den Stressmaßen vorliegt. Als nächsten Schritt könnten Sie die ordinalen Variablen "lösen", d. h. zulassen, dass äquivalente Werte der ursprünglichen Variablen verschiedene transformierte Werte erhalten. Beispiel: In der ersten Quelle betragen die Ähnlichkeiten zwischen *Tante* und *Sohn* und zwischen *Tante* und *Enkel* 85. Der "gebundene" Ansatz für ordinale Variablen erzwingt, dass die transformierten Werte dieser Ähnlichkeiten übereinstimmen, es gibt jedoch eigentlich keinen Grund für die Annahme, dass sie wirklich übereinstimmen sollten. In diesem Fall wird durch die Möglichkeit, die Ähnlichkeiten zu "lösen", eine unnötige Beschränkung aufgehoben.

Empfohlene Literatur

In folgenden Texten finden Sie weitere Informationen zur multidimensionalen Skalierung:

Commandeur, J. J. F., als auch W. J. Heiser. 1993. *Mathematical derivations in the proximity scaling (PROXSCAL) of symmetric data matrices*. Leiden: Department of Data Theory, Universität Leiden.

De Leeuw, J., als auch W. J. Heiser. 1980. Multidimensional scaling with restrictions on the configuration. In: *Multivariate Analysis, Vol. V,* P. R. Krishnaiah (Hg.). Amsterdam: North-Holland, 501–522.

Heiser, W. J. 1981. *Unfolding analysis of proximity data*. Leiden: Department of Data Theory, Universität Leiden.

Heiser, W. J., als auch F. M. T. A. Busing. 2004. Multidimensional scaling and unfolding of symmetric and asymmetric proximity relations. In: *Handbook of Quantitative Methodology for the Social Sciences,* D. Kaplan (Hg.). Thousand Oaks, Kalifornien: Sage Publications, Inc, 25–48.

Kruskal, J. B. 1964. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika*, 29, 1–28.

Kruskal, J. B. 1964. Nonmetric multidimensional scaling: A numerical method. *Psychometrika*, 29, 115–129.

Shepard, R. N. 1962. The analysis of proximities: Multidimensional scaling with an unknown distance function I. *Psychometrika*, 27, 125–140.

Shepard, R. N. 1962. The analysis of proximities: Multidimensional scaling with an unknown distance function II. *Psychometrika*, 27, 219–246.

Kapitel
15

Multidimensionale Entfaltung

In der Prozedur “Multidimensionale Entfaltung” wird versucht, eine gemeinsame, quantitative Skala zu finden, mit der Sie die Beziehung zwischen zwei Gruppen von Objekten visuell untersuchen können.

Beispiel: Bevorzugte Frühstücksartikel

In einer klassischen Studie (Green als auch Rao, 1972) wurden 21 MBA-Studenten der Wharton School mit ihren Lebensgefährten darum gebeten, 15 Frühstücksartikel in der Vorzugsreihenfolge von 1 = “am meisten bevorzugt” bis 15 = “am wenigsten bevorzugt” zu ordnen. Diese Informationen finden Sie in der Datei *breakfast_overall.sav*. Für weitere Informationen siehe Beispieldateien in Anhang A auf S. 326.

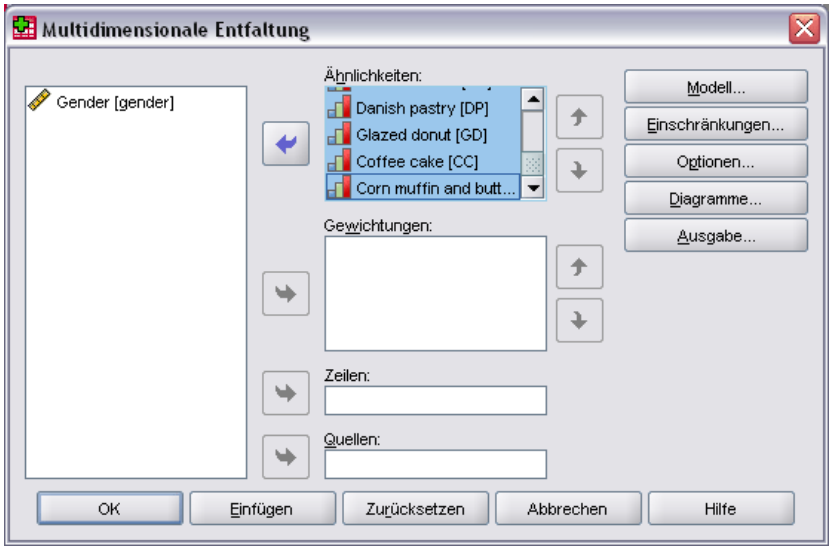
Die Ergebnisse der Studie stellen ein typisches Beispiel des in den meisten Algorithmen für die multidimensionale Entfaltung anzutreffenden Degenerationsproblems dar, das durch Penalisieren des Variationskoeffizienten für die transformierten Ähnlichkeiten gelöst wird (Busing, Groenen, als auch Heiser, 2005). Es wird eine degenerierte Lösung vorgestellt und erläutert, wie das Problem mit der multidimensionalen Entfaltung gelöst werden kann. Auf diese Weise können Sie die Unterscheidung der einzelnen Personen zwischen den Frühstücksartikeln ermitteln. Syntax, mit denen Sie diese Analysen nachvollziehen können, befindet sich in der Datei *prefscal_breakfast-overall.sps*.

Erstellen einer degenerierten Lösung

- Um eine Analyse vom Typ “Multidimensionale Entfaltung” durchzuführen, wählen Sie die folgenden Befehle aus den Menüs aus:

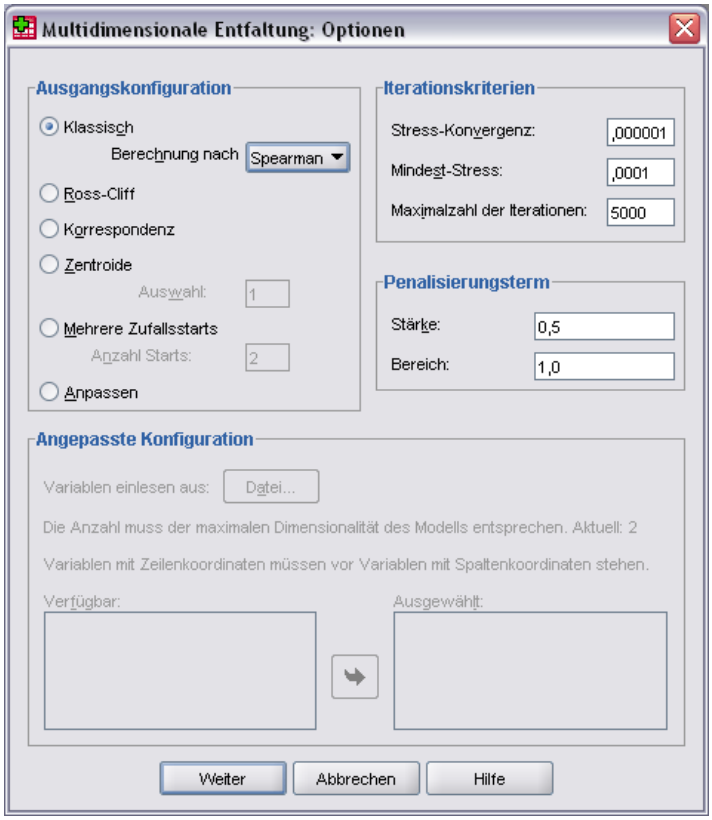
Analysieren
  Skalieren
    Multidimensionale Entfaltung...

Abbildung 15-1
Hauptdialogfeld "Multidimensionale Entfaltung"



- Wählen Sie *Toast pop-up* bis *Corn muffin and butter* als Ähnlichkeitsvariablen aus.
- Klicken Sie auf Optionen.

Abbildung 15-2
Dialogfeld "Optionen"



- Wählen Sie Spearman als Annahmemethode für die Ausgangskonfiguration "Klassisch" aus.
- Geben Sie im Gruppenfeld "Penalisierungsterm" 1,0 für "Stärke" und 0,0 für "Bereich" ein. Dadurch wird der Penalisierungsterm deaktiviert.
- Klicken Sie auf Weiter.
- Klicken Sie im Dialogfeld "Multidimensionale Entfaltung" auf OK.

Im Folgenden finden Sie die durch diese Optionen generierte Befehlssyntax:

```
PREFSCAL
  VARIABLES=TP BT EMM JD CT BMM HRB TMd BTJ TMn CB DP GD CC CMB
  /INITIAL=CLASSICAL (SPEARMAN)
  /TRANSFORMATION=NONE
  /PROXIMITIES=DISSIMILARITIES
  /CRITERIA=DIMENSIONS(2,2) DIFFSTRESS(.000001) MINSTRESS(.0001)
  MAXITER(5000)
  /PENALTY=LAMBDA(1.0) OMEGA(0.0)
  /PRINT=MEASURES COMMON
  /PLOT=COMMON .
```

- Mit dieser Syntax wird eine Analyse der Variablen *tp (Toast pop-up)* bis *cmb (Corn muffin and butter)* angegeben.

- Der Unterbefehl `INITIAL` gibt an, dass die Startwerte mithilfe der Spearman-Distanzen festgelegt werden sollen.
- Mit den angegebenen Werten für den Unterbefehl `PENALTY` wird im Wesentlichen die Penalisierung deaktiviert und somit das Kruskal-Stress-I minimiert. Dies führt zu einer degenerierten Lösung.
- Mit dem Unterbefehl `PLOT` werden Diagramme des gemeinsamen Raums angefordert.
- Für alle anderen Parameter gelten die Standardwerte.

Maße

Abbildung 15-3
Maße für degenerierte Lösung

Iterationen		154
Endgültiger Funktionswert		.0000990
Wertteile der Funktion	Stressteil	.0000990
	Penalisierungsteil	1.0000000
Fehlende Anpassung	Normalisierter Stress	.0000000
	Stress I nach Kruskal	.0000990
	Stress II nach Kruskal	.6129749
	S-Stress I nach Young	.0001980
	S-Stress II nach Young	.7703817
Güte der Anpassung	Erklärte Streuung	1.0000000
	Erklärte Varianz	.6230788
	Ermittelte Bevorzugungsordnungen	.7074830
	Spearman-Rho	.7450748
	Kendall-Tau-b	.6218729
Variationskoeffizienten	Variation der Ähnlichkeiten	.5590170
	Variation der transformierten Ähnlichkeiten	.0000924
	Variation der Distanzen	.1808765
Indizes für Degeneration	Quadratsumme von DeSarbos Indizes für Vermischung	117.3115413
	Shepards Index für Nicht-Degeneration	.0000000

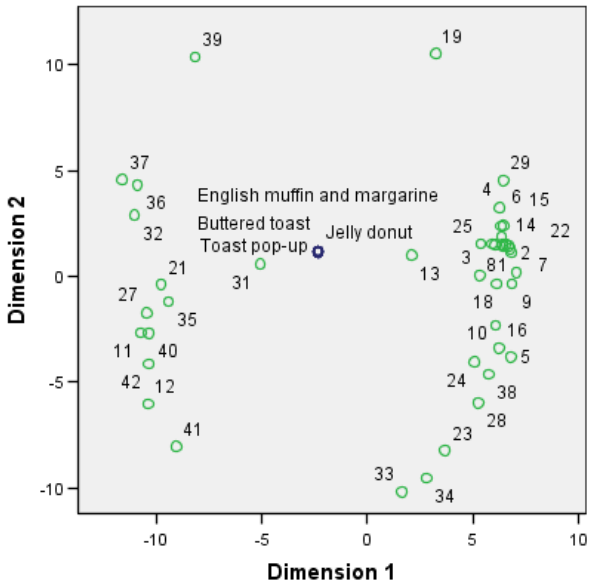
Der Algorithmus findet eine Lösung nach 154 Iterationen und konvergiert bei einem penalisierten Stress (markierten endgültigen Funktionswert) von 0,0000990. Da die Penalisierung deaktiviert wurde, entspricht der penalisierte Stress dem Kruskal-Stress-I (der Stressanteil des Funktionswerts entspricht dem Maß für die mangelhafte Güte der Anpassung nach Kruskal). Geringe Stresswerte deuten normalerweise darauf hin, dass die Lösung gut an die Daten angepasst ist. Es gibt jedoch mehrere Hinweise auf eine degenerierte Lösung:

- Der Variationskoeffizient für die transformierten Ähnlichkeiten ist relativ zum Variationskoeffizienten für die ursprünglichen Ähnlichkeiten sehr klein. Dies weist darauf hin, dass die transformierten Ähnlichkeiten für jede Zeile nahezu konstant sind und die Lösung somit keine Unterscheidung zwischen Objekten ermöglicht.

- Die Quadratsummen der Vermischungsindizes nach DeSarbo sind ein Maß für die Vermischung der Punkte in den unterschiedlichen Gruppen. Eine fehlende Vermischung deutet darauf hin, dass die Lösung möglicherweise degeneriert ist. Je näher sich der Wert bei 0 befindet, desto besser ist die Vermischung der Lösung. Der ausgegebene Wert ist sehr groß. Dies deutet darauf hin, dass die Lösung nicht vermischt ist.
- Shepards Index für Nicht-Degeneration, ein Prozentwert verschiedener Distanzen, ist gleich 0. Dies ist ein deutlicher Hinweis auf unzureichend unterschiedliche Abstände und die Lösung ist wahrscheinlich degeneriert.

Gemeinsamer Raum

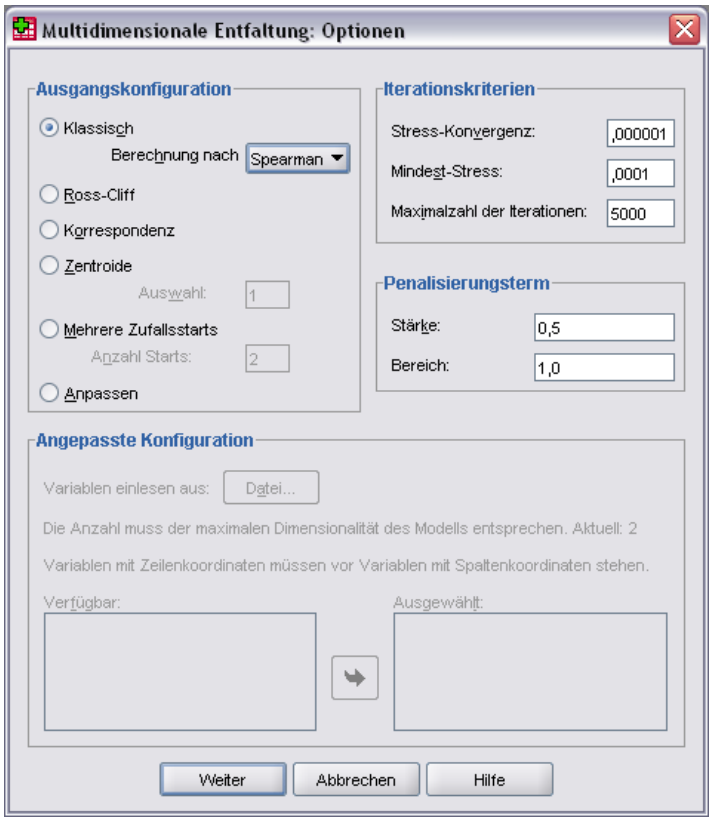
Abbildung 15-4
Verbundenes Diagramm des gemeinsamen Raums für eine degenerierte Lösung



Eine visuelle Bestätigung für die Degeneration der Lösung kann dem verbundenen Diagramm des gemeinsamen Raums von Zeilen- und Spaltenobjekten entnommen werden. Die Zeilenobjekte (Personen) befinden sich im Umfang eines Kreises mit den Spaltenobjekten (Frühstücksartikeln) als Mittelpunkt, deren Koordinaten in einem einzigen Punkt zusammenfallen.

Durchführen einer nicht degenerierten Analyse

Abbildung 15-5
Dialogfeld "Optionen"



- Um eine nicht degenerierte Lösung zu erstellen, klicken Sie auf die Schaltfläche "Zuletzt verwendete Dialogfelder" und wählen Sie Multidimensionale Entfaltung aus.
- Klicken Sie im Dialogfeld "Multidimensionale Entfaltung" auf Optionen.
- Geben Sie im Gruppenfeld "Penalisierungsterm" 0,5 für "Stärke" und 1,0 für "Bereich" ein. Dadurch wird der Penalisierungsterm deaktiviert.
- Klicken Sie auf Weiter.
- Klicken Sie im Dialogfeld "Multidimensionale Entfaltung" auf OK.

Im Folgenden finden Sie die durch diese Optionen generierte Befehlssyntax:

```
PREFSCAL
  VARIABLES=TP BT EMM JD CT BMM HRB TMd BTJ TMn CB DP GD CC CMB
  /INITIAL=CLASSICAL (SPEARMAN)
  /TRANSFORMATION=NONE
  /PROXIMITIES=DISSIMILARITIES
  /CRITERIA=DIMENSIONS(2,2) DIFFSTRESS(.000001) MINSTRESS(.0001)
  MAXITER(5000)
  /PENALTY=LAMBDA(0.5) OMEGA(1.0)
  /PRINT=MEASURES COMMON
  /PLOT=COMMON .
```

- Die einzige Änderung besteht im Unterbefehl PENALTY. LAMBDA wurde auf 0,5 und OMEGA auf 1,0 (die Standardwerte) festgelegt.

Maße

Abbildung 15-6
Maße für nicht degenerierte Lösung

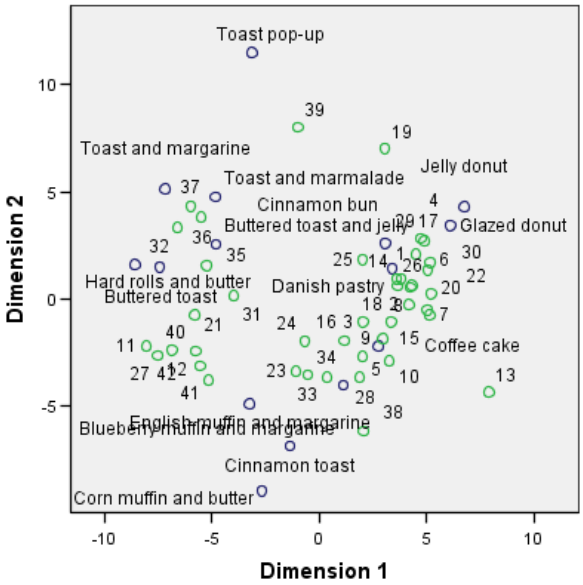
Iterationen		157
Endgültiger Funktionswert		,6848930
Wertteile der Funktion	Stressteil	,2428268
	Penalisierungsteil	1,9317409
Fehlende Anpassung	Normalisierter Stress	,0583589
	Stress I nach Kruskal	,2415758
	Stress II nach Kruskal	,5875599
	S-Stress I nach Young	,3446361
	S-Stress II nach Young	,5030127
Güte der Anpassung	Erklärte Streuung	,9416411
	Erklärte Varianz	,7651552
	Ermittelte Bevorzugungsordnungen	,7818594
	Spearman-Rho	,8179181
	Kendall-Tau-b	,6916725
Variationskoeffizienten	Variation der Ähnlichkeiten	,5590170
	Variation der transformierten Ähnlichkeiten	,6006156
	Variation der Distanzen	,4833617
Indizes für Degeneration	Quadratsumme von DeSarbos Indizes für Vermischung	,1590979
	Shepards Index für Nicht-Degeneration	,7895692

Die bei den Maßen für die degenerierte Lösung beobachteten Probleme wurden hier behoben.

- Der normalisierte Stress beträgt nicht mehr 0.
- Der Variationskoeffizient für die transformierten Ähnlichkeiten weist nun einen ähnlichen Wert wie der Variationskoeffizient für die ursprünglichen Ähnlichkeiten auf.
- Die Vermischungsindizes nach DeSarbo liegen wesentlich näher bei 0 und deuten somit auf eine erheblich bessere Vermischung hin.
- Shepards Index für Nicht-Degeneration, ein Prozentwert verschiedener Distanzen, liegt jetzt bei nahezu 80 %. Die Distanzen sind ausreichend, und die Lösung ist wahrscheinlich nicht degeneriert.

Gemeinsamer Raum

Abbildung 15-7
Verbundenes Diagramm des gemeinsamen Raums für eine nicht degenerierte Lösung



Das verbundene Diagramm des gemeinsamen Raums ermöglicht eine Interpretation der Dimensionen. Die horizontale Dimension scheint eine Unterscheidung zwischen weichem und harten Brot oder Toast aufzuweisen, wobei die Artikel nach rechts entlang der Achse weicher werden. Bei der vertikalen Dimension ist keine eindeutige Interpretation möglich, jedoch ist möglicherweise eine Unterscheidung nach Üblichkeit vorhanden, bei der die Artikel nach unten entlang der Achse "konventioneller" werden.

Es entstehen mehrere Gruppen von Frühstücksartikeln. Zum Beispiel bilden Donuts, Cinnamon Buns und Danish Pastry eine Gruppe von weichen und eher unkonventionellen Artikeln. Muffins und Cinnamon Toast bilden eine Gruppe von härteren, jedoch auch "konventionelleren" Artikeln. Die anderen Toasts und härteren Brötchen bilden eine Gruppe von harten und eher unkonventionellen Artikeln. "Toast pop-up" ist ein harter und äußerst unkonventioneller Artikel.

Die durch die Zeilenobjekte dargestellten Personen sind deutlich nach der Bevorzugung von harten oder weichen Artikeln gruppiert und weisen in der vertikalen Dimension starke Variationen innerhalb der Gruppen auf.

Beispiel: Dreifach-Entfaltung von bevorzugten Frühstücksartikeln

In einer klassischen Studie (Green et al., 1972) wurden 21 MBA-Studenten der Wharton School mit ihren Lebensgefährten darum gebeten, 15 Frühstücksartikel in der Vorzugsreihenfolge von 1 = "am meisten bevorzugt" bis 15 = "am wenigsten bevorzugt" zu ordnen. Die Bevorzugungen wurden in sechs unterschiedlichen Szenarien erfasst, von "Overall preference" (Allgemein bevorzugt) bis "Snack, with beverage only" (Imbiss, nur mit Getränk). Diese Informationen finden Sie in der Datei *breakfast.sav*. Für weitere Informationen siehe Beispieldateien in Anhang A auf S. 326.

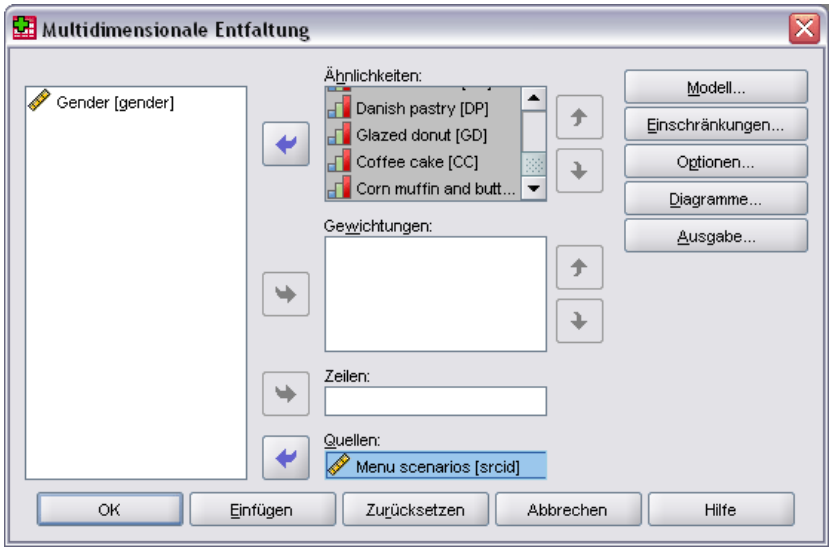
Die sechs Szenarien können als getrennte Quellen behandelt werden. Führen Sie mit PREFSCAL eine Drei-fach-Entfaltung der Zeilen, Spalten und Quellen durch. Syntax, mit denen Sie diese Analysen nachvollziehen können, befindet sich in der Datei *prefscal_breakfast.sps*.

Durchführung der Analyse

- Um eine Analyse vom Tyü "Multidimensionale Entfaltung" durchzuführen, wählen Sie die folgenden Befehle aus den Menüs aus:

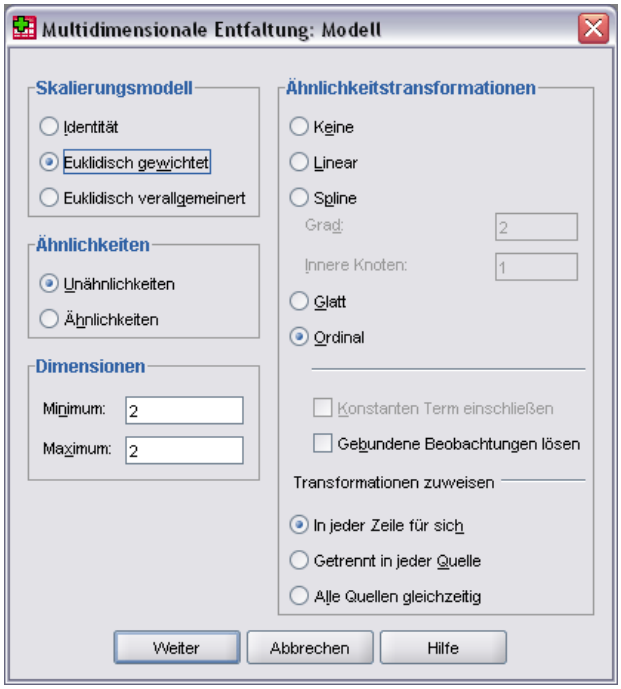
Analysieren
  Skalieren
    Multidimensionale Entfaltung...

Abbildung 15-8
Hauptdialogfeld "Multidimensionale Entfaltung"



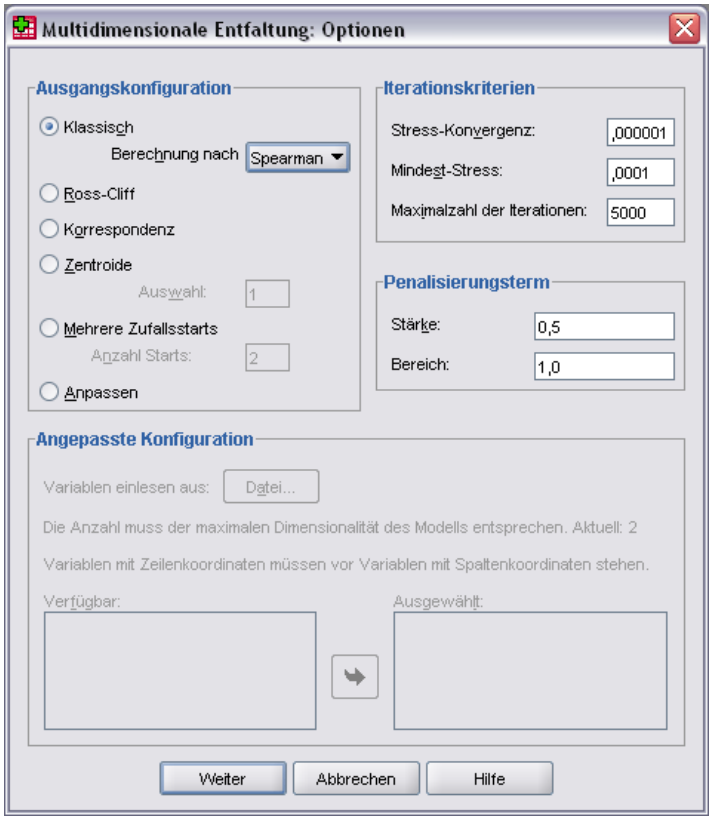
- Wählen Sie *Toast pop-up* bis *Corn muffin and butter* als Ähnlichkeitsvariablen aus.
- Wählen Sie *Menu scenarios* als Ursprungsvariable.
- Klicken Sie auf Modell.

Abbildung 15-9
Dialogfeld "Modell"



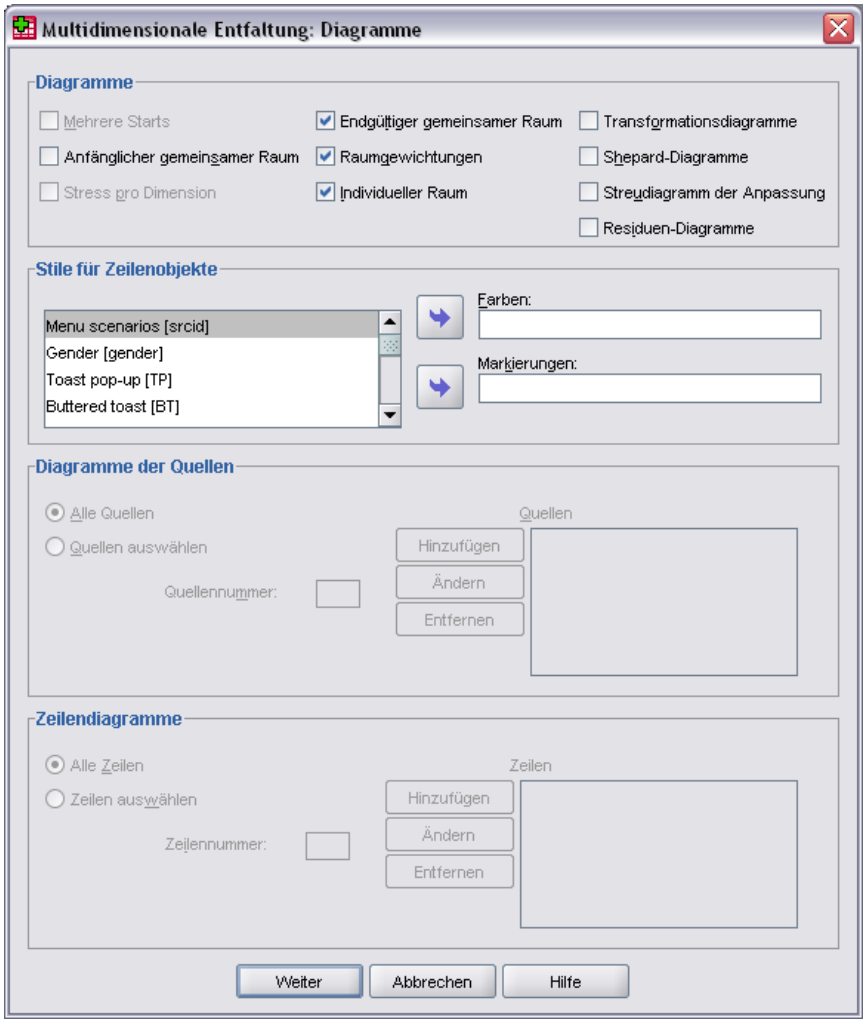
- Wählen Sie Euklidisch gewichtet als Skalierungsmodell.
- Klicken Sie auf Weiter.
- Klicken Sie im Dialogfeld "Multidimensionale Entfaltung" auf Optionen.

Abbildung 15-10
Dialogfeld "Optionen"



- Wählen Sie Spearman als Annahmemethode für die Ausgangskonfiguration "Klassisch" aus.
- Klicken Sie auf Weiter.
- Klicken Sie im Dialogfeld "Multidimensionale Entfaltung" auf Diagramme.

Abbildung 15-11
Dialogfeld "Diagramme"



- Wählen Sie in der Gruppe "Diagramme" die Option Individueller Raum.
- Klicken Sie auf Weiter.
- Klicken Sie im Dialogfeld "Multidimensionale Entfaltung" auf OK.

Im Folgenden finden Sie die durch diese Optionen generierte Befehlssyntax:

```
PREFSCAL
  VARIABLES=TP BT EMM JD CT BMM HRB TMd BTJ TMn CB DP GD CC CMB
  /INPUT=SOURCES(srcid )
  /INITIAL=CLASSICAL (SPEARMAN)
  /CONDITION=ROW
  /TRANSFORMATION=NONE
  /PROXIMITIES=DISSIMILARITIES
  /MODEL=WEIGHTED
  /CRITERIA=DIMENSIONS(2,2) DIFFSTRESS(.000001) MINSTRESS(.0001)
  MAXITER(5000)
  /PENALTY=LAMBDA(0.5) OMEGA(1.0)
```

```
/PRINT=MEASURES COMMON
/PLOT=COMMON WEIGHTS INDIVIDUAL ( ALL ) .
```

- Mit dieser Syntax wird eine Analyse der Variablen *tp (Toast pop-up)* bis *cmb (Corn muffin and butter)* angegeben. Mit der Variablen *srcid* werden die Quellen identifiziert.
- Der Unterbefehl `INITIAL` gibt an, dass die Startwerte mithilfe der Spearman-Distanzen festgelegt werden sollen.
- Der Unterbefehl `MODEL` gibt ein gewichtetes euklidisches Modell an, bei dem in jedem individuellen Raum die Dimensionen des gemeinsamen Raums unterschiedlich gewichtet werden können.
- Mit dem Unterbefehl `PLOT` werden Diagramme des gemeinsamen Raums, individuelle Räume und individuelle Raumgewichtungen angefordert.
- Für alle anderen Parameter gelten die Standardwerte.

Maße

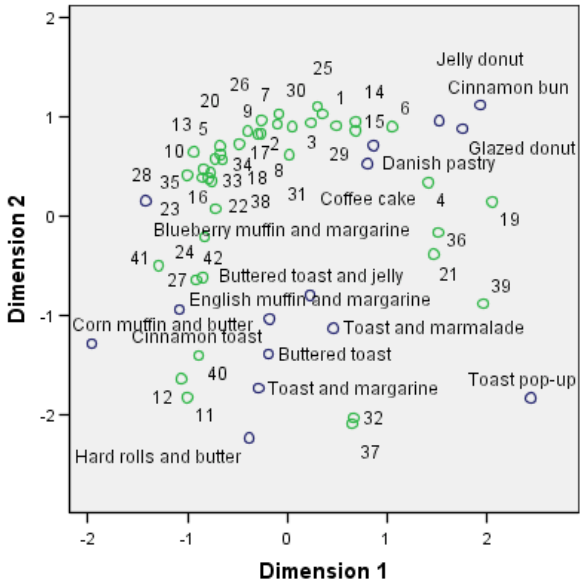
Abbildung 15-12
Maße

Iterationen		481
Endgültiger Funktionswert		.8199642
Wertteile der Funktion	Stressteil	.3680994
	Penalisierungsteil	1.8265211
Fehlende Anpassung	Normalisierter Stress	.1335343
	Stress I nach Kruskal	.3654234
	Stress II nach Kruskal	.9780824
	S-Stress I nach Young	.4938016
	S-Stress II nach Young	.6912352
Güte der Anpassung	Erklärte Streuung	.8664657
	Erklärte Varianz	.5024853
	Ermittelte Bevorzugungsordnungen	.7025321
	Spearman-Rho	.6271702
	Kendall-Tau-b	.4991188
Variationskoeffizienten	Variation der Ähnlichkeiten	.5590170
	Variation der transformierten Ähnlichkeiten	.6378878
	Variation der Distanzen	.4484515
Indizes für Degeneration	Quadratsumme von DeSarbos Indizes für Vermischung	.2199287
	Shepards Index für Nicht-Degeneration	.7643613

Der Algorithmus konvergiert nach 481 Iterationen bei einem endgültigen penalisierten Stress von 0,8199642. Die Variationskoeffizienten und Shepards Index sind ausreichend groß und die Indizes nach DeSarbo ausreichend klein, um den Schluss zuzulassen, dass keine Degenerationsprobleme vorliegen.

Gemeinsamer Raum

Abbildung 15-13
Verbundenes Diagramm des gemeinsamen Raums



Das verbundene Diagramm des gemeinsamen Raums zeigt eine endgültige Konfiguration, die der Zwei-fach-Analyse der allgemeinen Bevorzugung stark ähnelt, wobei die Lösung an der 45-Grad-Achse gespiegelt ist. Daher scheint nun die vertikale Dimension eine Unterscheidung zwischen weichem und harten Brot oder Toast aufzuweisen, wobei die Artikel nach oben entlang der Achse weicher werden. Bei der horizontalen Dimension ist nun keine eindeutige Interpretation möglich, jedoch ist möglicherweise eine Unterscheidung nach Bequemlichkeit vorhanden, bei der die Artikel nach links entlang der Achse “konventioneller” werden.

Die durch die Zeilenobjekte dargestellten Personen sind immer noch deutlich nach der Bevorzugung von harten oder weichen Artikeln gruppiert und weisen in der horizontalen Dimension starke Variationen innerhalb der Gruppen auf.

Individueller Raum

Abbildung 15-14
Dimensionsgewichtungen

		Dimension		
		1	2	Spezifität[a]
Quelle	Overall preference	3,235	4,297	,186
	Breakfast, with juice, bacon and eggs, and beverage	4,883	2,193	,457
	Breakfast, with juice, cold cereal, and beverage	4,131	3,438	,109
	Breakfast, with juice, pancakes, sausage, and beverage	4,291	3,267	,164
	Breakfast, with beverage only	3,124	4,413	,223
	Snack, with beverage only	2,750	4,541	,313
Wichtigkeit[b]		,504	,496	

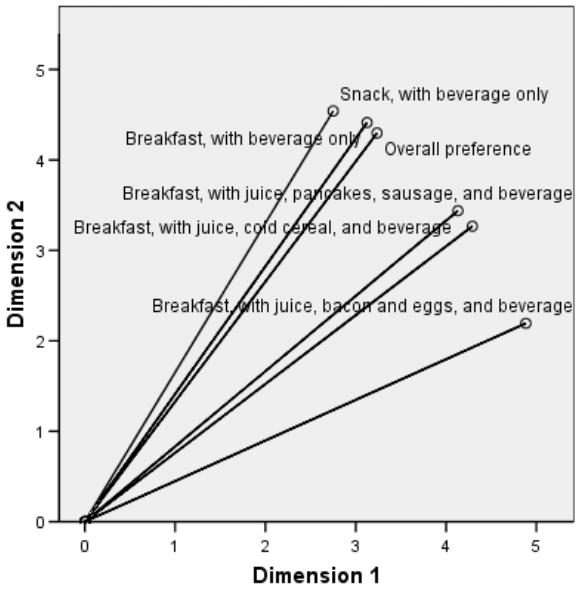
a. Die Spezifität gibt an, wie typisch eine Quelle ist. Der Bereich für die Spezifität liegt zwischen null und eins. Dabei weist null auf eine durchschnittliche Quelle mit identischen Dimensionsgewichtungen hin. Eins weist auf eine sehr spezifische Quelle mit einer ungewöhnlich großen Dimensionsgewichtung und anderen Gewichtungen nahe null hin.

b. Relative Wichtigkeit jeder Dimension als Verhältnis der Quadratsumme einer Dimension zur Gesamtquadratsumme.

Für jede Quelle wird ein individueller Raum berechnet. In den Dimensionsgewichtungen wird die Beteiligung der individuellen Räume an den Dimensionen des gemeinsamen Raums gezeigt. Eine größere Gewichtung weist auf eine größere Distanz im individuellen Raum und somit auf eine größere Unterscheidung zwischen den Objekten in dieser Dimension für diesen individuellen Raum hin.

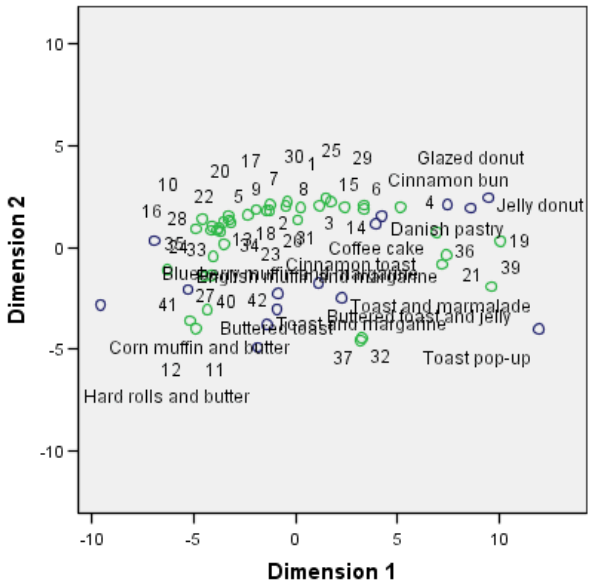
- **Spezifität** ist ein Maß der Abweichung eines individuellen Raums vom gemeinsamen Raum. Ein individueller Raum, der mit dem gemeinsamen Raum identisch ist, hat identische Dimensionsgewichtungen und eine Spezifität von 0. Ein individueller Raum, der in einer bestimmten Dimension einzigartig ist, weist hingegen eine einzige große Dimensionsgewichtung und eine Spezifität von 1 auf. In diesem Fall sind die Quellen mit den größten Abweichungen *Breakfast, with juice, bacon and eggs, and beverage* sowie *Snack, with beverage only*.
- **Wichtigkeit** ist ein Maß für den relativen Anteil jeder Dimension an der Lösung. In diesem Fall sind die Dimensionen gleich wichtig.

Abbildung 15-15
Dimensionsgewichtungen



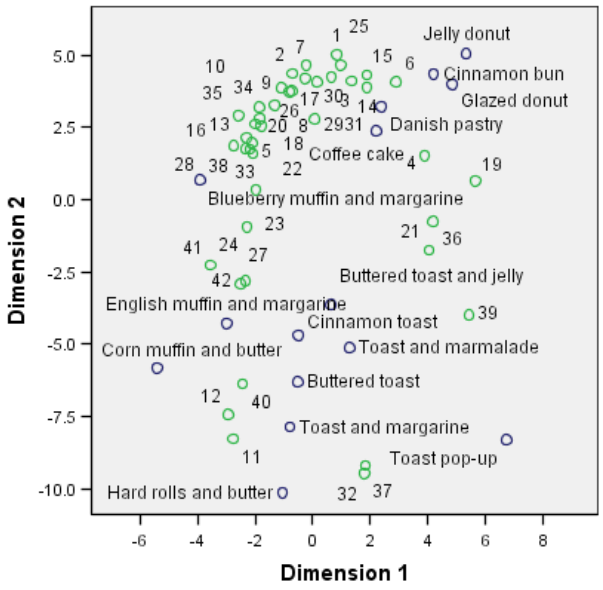
Das Diagramm der Dimensionsgewichtungen bietet eine Visualisierung der Gewichtungstabelle. *Breakfast, with juice, bacon and eggs, and beverage* und *Snack, with beverage only* befinden sich am nächsten an den Dimensionsachsen, sind jedoch beide nicht besonders spezifisch in einer bestimmten Dimension.

Abbildung 15-16
Verbundenes Diagramm des individuellen Raums "Breakfast, with juice, bacon and eggs, and beverage"



Im verbundenen Diagramm des individuellen Raums *Breakfast, with juice, bacon and eggs, and beverage* sind die Auswirkungen dieses Szenarios auf die Bevorzugungen dargestellt. Diese Quelle ist stärker an der ersten Dimension beteiligt, sodass die Unterscheidung zwischen den Artikeln meist auf die erste Dimension zurückzuführen ist.

Abbildung 15-17
Verbundenes Diagramm des individuellen Raums "Snack, with beverage only"



Im verbundenen Diagramm des individuellen Raums *Snack, with beverage only* sind die Auswirkungen dieses Szenarios auf die Bevorzugungen dargestellt. Diese Quelle ist stärker an der zweiten Dimension beteiligt, sodass die Unterscheidung zwischen den Artikeln meist auf die zweite Dimension zurückzuführen ist. Aufgrund der geringen Spezifität dieser Quelle ist entlang der ersten Dimension jedoch immer noch eine deutliche Unterscheidung vorhanden.

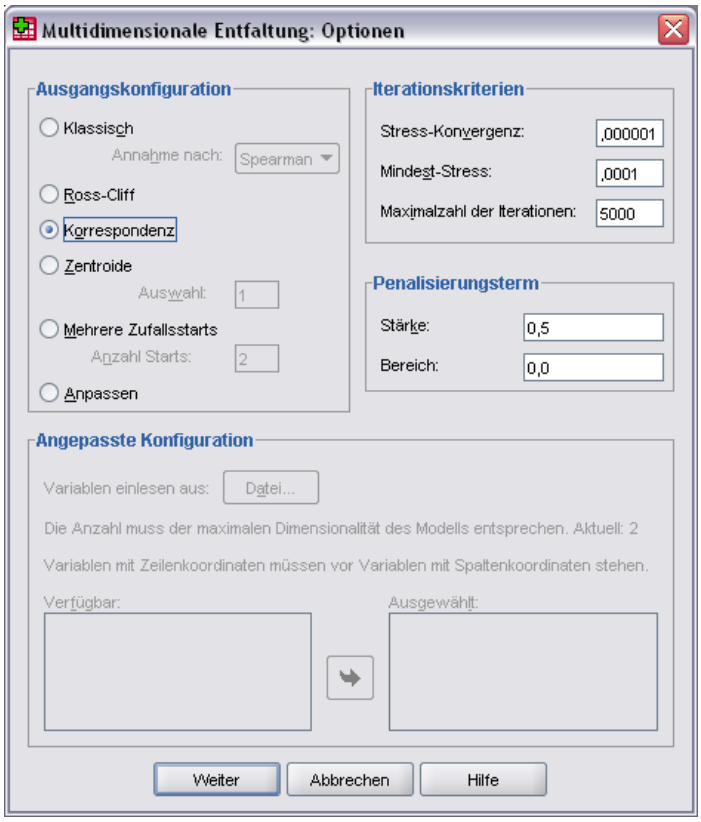
Verwenden einer anderen Ausgangskonfiguration

Die endgültige Konfiguration kann von den Startpunkten abhängen, die dem Algorithmus übergeben werden. Im Idealfall sollte die allgemeine Struktur der Lösung gleich bleiben, andernfalls fällt das Ermitteln der richtigen Lösung möglicherweise schwer. Durch die Verwendung verschiedener Ausgangskonfigurationen, z. B. mit einem Korrespondenzstart der Drei-fach-Analyse der Frühstücksdaten, können jedoch Einzelheiten stärker herausgestellt werden.

- Um eine Lösung mit einem Korrespondenzstart zu erstellen, klicken Sie auf die Schaltfläche "Zuletzt verwendete Dialogfelder" und wählen Sie Multidimensionale Entfaltung aus.

- Klicken Sie im Dialogfeld “Multidimensionale Entfaltung” auf Optionen.

Abbildung 15-18
Dialogfeld “Optionen”



- Wählen Sie im Gruppenfeld “Ausgangskonfiguration” die Option Korrespondenz.
- Klicken Sie auf Weiter.
- Klicken Sie im Dialogfeld “Multidimensionale Entfaltung” auf OK.

Im Folgenden finden Sie die durch diese Optionen generierte Befehlssyntax:

```
PREFSCAL
  VARIABLES=TP BT EMM JD CT BMM HRB TMd BTJ TMn CB DP GD CC CMB
  /INPUT=SOURCES(srcid )
  /INITIAL=CORRESPONDENCE
  /TRANSFORMATION=NONE
  /PROXIMITIES=DISSIMILARITIES
  /CRITERIA=DIMENSIONS(2,2) DIFFSTRESS(.000001) MINSTRESS(.0001)
  MAXITER(5000)
  /PENALTY=LAMBDA(0.5) OMEGA(1.0)
  /PRINT=MEASURES COMMON
  /PLOT=COMMON WEIGHTS INDIVIDUAL ( ALL ) .
```

- Die einzige Änderung besteht im Unterbefehl `INITIAL`. Die Ausgangskonfiguration wurde als `CORRESPONDENCE` festgelegt. Dabei werden die Ergebnisse einer Korrespondenzanalyse der invertierten Daten (Ähnlichkeiten anstelle von Unähnlichkeiten) mit symmetrischer Normalisierung der Zeilen- und Spaltenwerte verwendet.

Maße

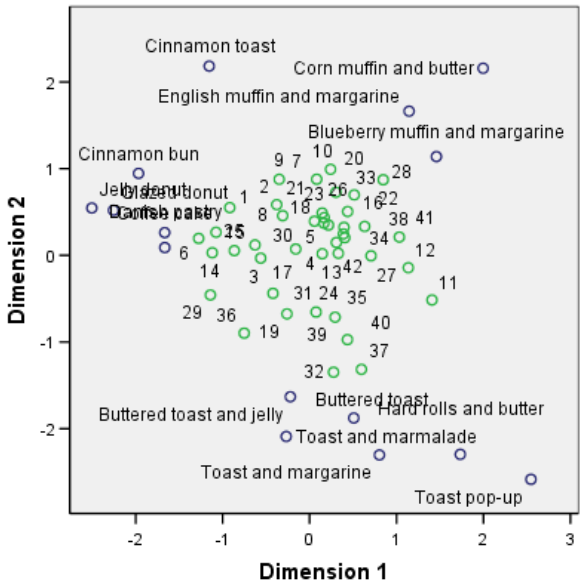
Abbildung 15-19
Maße für Ausgangskonfiguration mit Korrespondenzanalyse

Iterationen		385
Endgültiger Funktionswert		.8140741
Wertteile der Funktion	Stressteil	.3493640
	Penalisierungsteil	1.8969229
Fehlende Anpassung	Normalisierter Stress	.1212145
	Stress I nach Kruskal	.3481587
	Stress II nach Kruskal	1.0770522
	S-Stress I nach Young	.4812632
	S-Stress II nach Young	.6871733
Güte der Anpassung	Erklärte Streuung	.8787855
	Erklärte Varianz	.5183498
	Ermittelte Bevorzugungsordnungen	.7174981
	Spearman-Rho	.6446272
	Kendall-Tau-b	.5165230
Variationskoeffizienten	Variation der Ähnlichkeiten	.5590170
	Variation der transformierten Ähnlichkeiten	.6122308
	Variation der Distanzen	.4043695
Indizes für Degeneration	Quadratsumme von DeSarbos Indizes für Vermischung	1.7571887
	Shepards Index für Nicht-Degeneration	.7532124

Der Algorithmus konvergiert nach 385 Iterationen bei einem endgültigen penalisierten Stress von 0,8140741. Diese Statistik, die mangelhafte Güte der Anpassung, die Güte der Anpassung, die Variationskoeffizienten und Shepards Index ähneln stark der Lösung mit dem klassischen Spearman-Start. DeSarbos Indizes weichen mit einem Wert von 1,7571887 gegenüber 0,2199287 ab. Dies deutet darauf hin, dass die Lösung mit dem Korrespondenzstart nicht gleich gut vermischt ist. Sie können dem verbundenen Diagramm des gemeinsamen Raums entnehmen, wie sich dies auf die Lösung auswirkt.

Gemeinsamer Raum

Abbildung 15-20
Verbundenes Diagramm des gemeinsamen Raums für Ausgangskonfiguration mit Korrespondenzanalyse



Das gemeinsame Diagramm des gemeinsamen Raums weist eine endgültige Konfiguration auf, die der Analyse mit klassischer Spearman-Ausgangskonfiguration ähnelt. Die Spaltenobjekte (Frühstücksartikel) befinden sich jedoch im Umkreis der Zeilenobjekte (Personen), anstatt mit diesen vermischt zu sein.

Individueller Raum

Abbildung 15-21
Dimensionsgewichtungen für Ausgangskonfiguration mit Korrespondenzanalyse

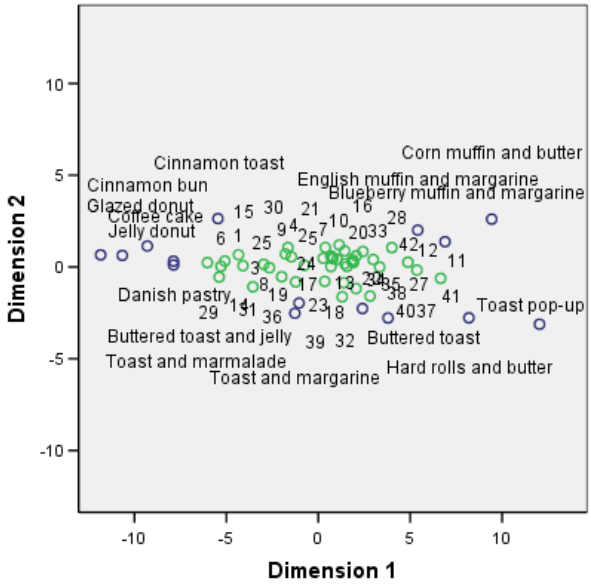
		Dimension		
		1	2	Spezifität[a]
Quelle	Overall preference	2,836	3,877	,279
	Breakfast, with juice, bacon and eggs, and beverage	4,727	1,207	,636
	Breakfast, with juice, cold cereal, and beverage	4,183	2,377	,263
	Breakfast, with juice, pancakes, sausage, and beverage	4,412	1,993	,389
	Breakfast, with beverage only	2,605	4,050	,351
	Snack, with beverage only	1,864	4,415	,552
Wichtigkeit[b]		,556	,444	

a. Die Spezifität gibt an, wie typisch eine Quelle ist. Der Bereich für die Spezifität liegt zwischen null und eins. Dabei weist null auf eine durchschnittliche Quelle mit identischen Dimensionsgewichtungen hin. Eins weist auf eine sehr spezifische Quelle mit einer ungewöhnlich großen Dimensionsgewichtung und anderen Gewichtungen nahe null hin.

b. Relative Wichtigkeit jeder Dimension als Verhältnis der Quadratsumme einer Dimension zur Gesamtquadratsumme.

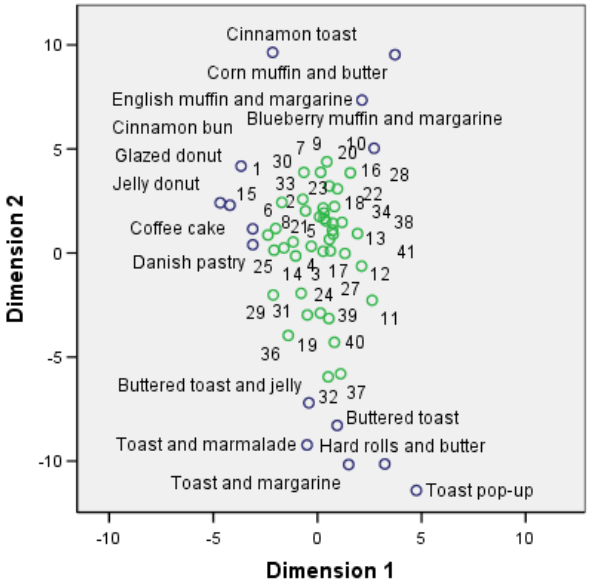
In der Ausgangskonfiguration mit Korrespondenzanalyse weist jeder individuelle Raum eine höhere Spezifität auf, d. h. die Situationen, in denen die Teilnehmer den Frühstücksartikeln ordneten, sind stärker einer bestimmten Dimension zugeordnet. Die Quellen mit den größten Abweichungen bleiben *Breakfast, with juice, bacon and eggs, and beverage* und *Snack, with beverage only*.

Abbildung 15-22
Verbundenes Diagramm des individuellen Raums "Breakfast, with juice, bacon and eggs, and beverage" für Ausgangskonfiguration mit Korrespondenzanalyse



Die höhere Spezifität ist im verbundenen Diagramm des individuellen Raums *Breakfast, with juice, bacon and eggs, and beverage* klar zu erkennen. Die Quelle ist noch stärker an der ersten Dimension als beim klassischen Spearman-Start beteiligt, sodass die Zeilen- und Spaltenobjekte auf der vertikalen Achse etwas weniger und auf der horizontalen Achse etwas mehr Variation aufweisen.

Abbildung 15-23
Verbundenes Diagramm des individuellen Raums "Snack, with beverage only" für Ausgangskonfiguration mit Korrespondenzanalyse



Im verbundenen Diagramm des individuellen Raums *Snack, with beverage only* ist zu erkennen, dass die Zeilen- und Spaltenobjekte dichter an einer vertikalen Linie liegen als beim klassischen Spearman-Start.

Beispiel: Untersuchen der Angemessenheit von Verhaltensweisen in bestimmten Situationen

In einem klassischen Beispiel (Price als auch Bouffard, 1974) wurden 52 Schüler/Studenten gebeten, die Kombinationen aus 15 Situationen und 15 Verhaltensweisen auf einer 10-Punkte-Skala von 0 = "ausgesprochen angemessen" bis 9 = "ausgesprochen unangemessen" zu bewerten. Die Werte werden über die einzelnen Personen gemittelt und als Unähnlichkeiten verwendet.

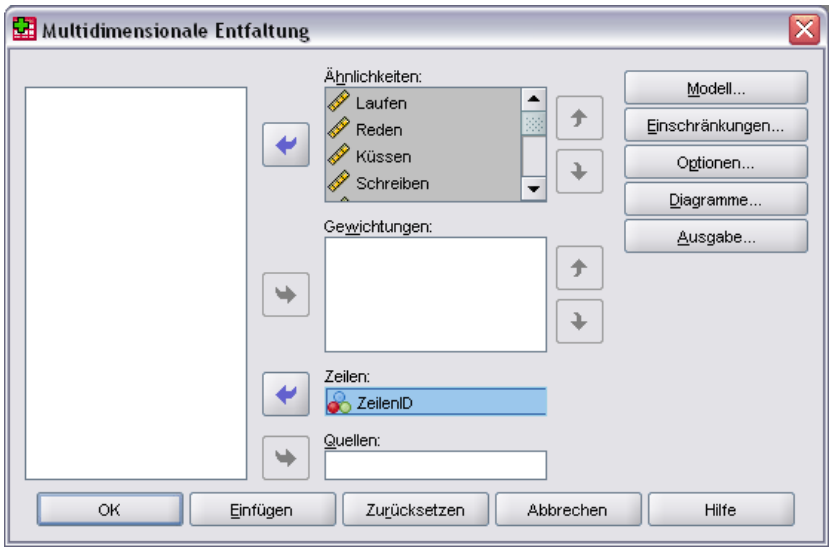
Diese Informationen finden Sie in der Datei *behavior.sav*. Für weitere Informationen siehe Beispieldateien in Anhang A auf S. 326. Suchen Sie mithilfe der multidimensionalen Entfaltung Gruppen von ähnlichen Situationen sowie die Verhaltensweisen, mit denen sie am ehesten in Verbindung gebracht werden. Syntax, mit denen Sie diese Analysen nachvollziehen können, befindet sich in der Datei *prefscal_behavior.sps*.

Durchführung der Analyse

- Um eine Analyse vom Tyü “Multidimensionale Entfaltung” durchzuführen, wählen Sie die folgenden Befehle aus den Menüs aus:

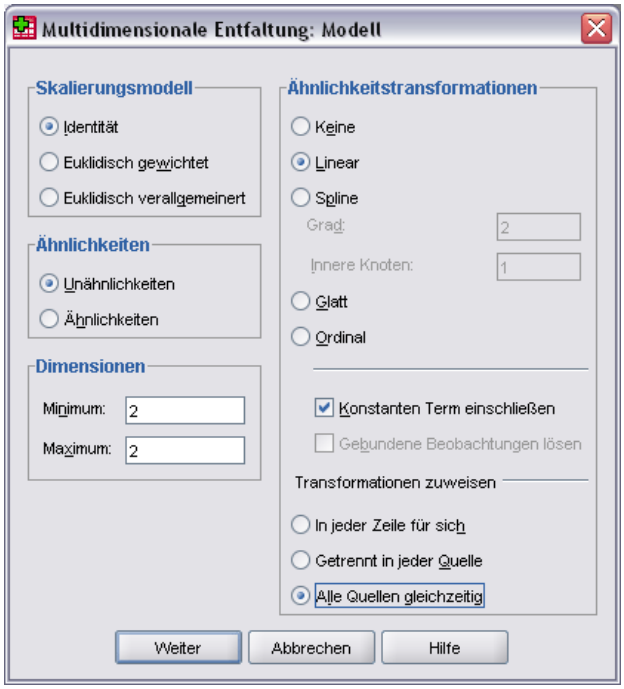
  Analysieren
  Skalieren
  Multidimensionale Entfaltung...

Abbildung 15-24
Hauptdialogfeld “Multidimensionale Entfaltung”



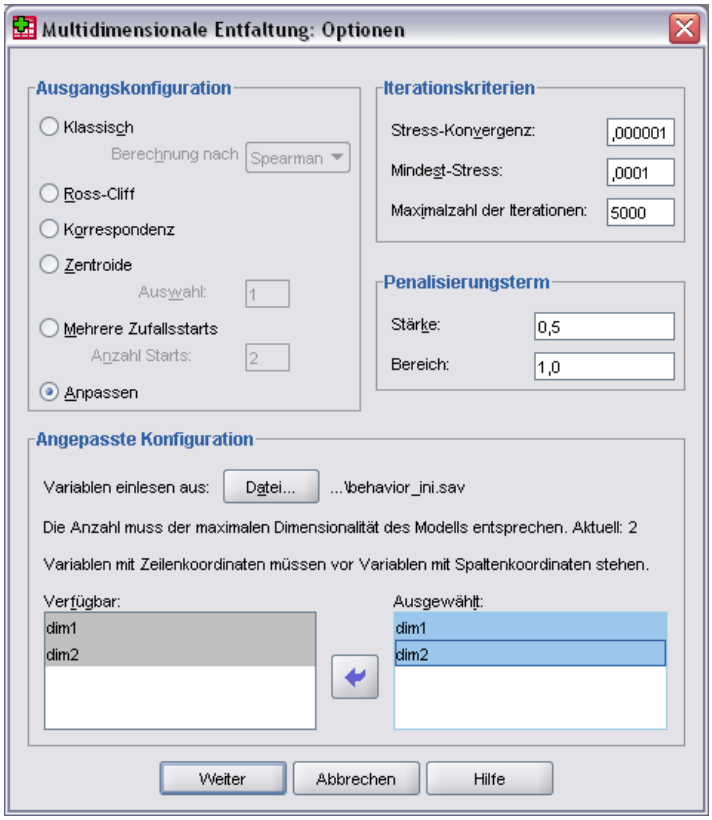
- Wählen Sie *Laufen* bis *Schreien* als Ähnlichkeitsvariablen aus.
- Wählen Sie *ZeilenID* als Zeilenvariable aus.
- Klicken Sie auf Modell.

Abbildung 15-25
Dialogfeld "Modell"



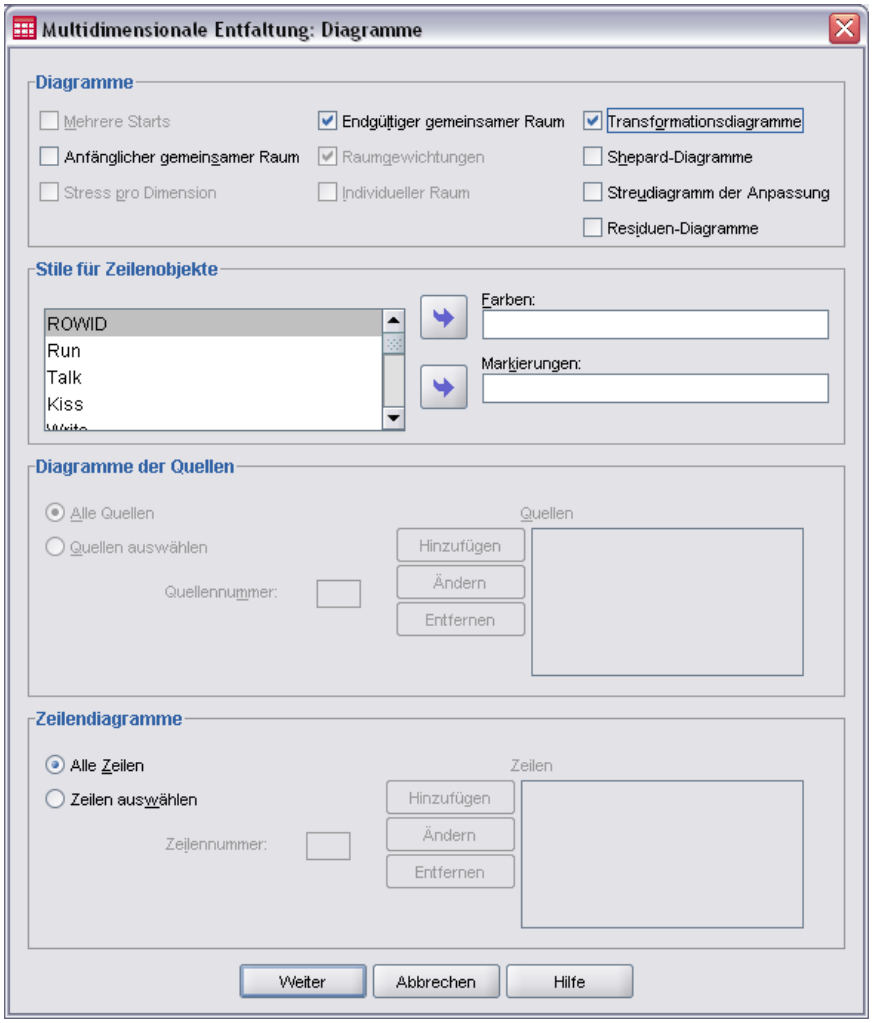
- Wählen Sie Linear als Ähnlichkeitstransformation aus und aktivieren Sie die Option Konstanten Term einschließen.
- Wählen Sie unter "Transformationen zuweisen" die Option Alle Quellen gleichzeitig aus.
- Klicken Sie auf Weiter.
- Klicken Sie im Dialogfeld "Multidimensionale Entfaltung" auf Optionen.

Abbildung 15-26
Dialogfeld "Optionen"



- Wählen Sie im Gruppenfeld "Ausgangskonfiguration" die Option Anpassen.
- Wechseln Sie zu *behavior_ini.sav* und wählen Sie diese Datei als die Datei aus, die die benutzerdefinierte Ausgangskonfiguration enthält. Für weitere Informationen siehe Beispieldateien in Anhang A auf S. 326.
- Wählen Sie *dim1* und *dim2* als Variablen für die Ausgangskonfiguration aus.
- Klicken Sie auf Weiter.
- Klicken Sie im Dialogfeld "Multidimensionale Entfaltung" auf Diagramme.

Abbildung 15-27
Dialogfeld "Diagramme"



- Wählen Sie in der Gruppe "Diagramme" die Option Transformationsdiagramme.
- Klicken Sie auf Weiter.
- Klicken Sie im Dialogfeld "Multidimensionale Entfaltung" auf OK.

Im Folgenden finden Sie die durch diese Optionen generierte Befehlssyntax:

```
PREFSCAL
  VARIABLES=Laufen Reden Küssen Schreiben Essen Schlafen Murmeln Lesen
  Kämpfen Rülpsen Diskutieren Springen Weinen Lachen Schreien
  /INPUT=ROWS(ROWID )
  /INITIAL=( 'samplesDirectory/behavior_ini.sav' )
  dim1 dim2
  /CONDITION=UNCONDITIONAL
  /TRANSFORMATION=LINEAR (INTERCEPT)
  /PROXIMITIES=DISSIMILARITIES
  /MODEL=IDENTITY
  /CRITERIA=DIMENSIONS(2,2) DIFFSTRESS(.000001) MINSTRESS(.0001)
```

```
MAXITER(5000)
/PENALTY=LAMBDA(0.5) OMEGA(1.0)
/PRINT=MEASURES COMMON
/PLOT=COMMON TRANSFORMATIONS .
```

- Mit dieser Syntax wird eine Analyse der Variablen *run* bis *shout* angegeben. Mit der Variablen *rowid* werden die Zeilen identifiziert.
- Der Unterbefehl `INITIAL` gibt an, dass die Startwerte aus der Datei *behavior_ini.sav* entnommen werden sollen. Die Zeilen- und Spaltenkoordinaten sind gestapelt, wobei die Spaltenkoordinaten auf die Zeilenkoordinaten folgen.
- Der Unterbefehl `CONDITION` gibt an, dass alle Ähnlichkeiten miteinander verglichen werden können. Dies trifft auf die vorliegende Analyse zu, da es möglich sein sollte, die Ähnlichkeiten zwischen dem Laufen in einem Park und dem Laufen in einer Kirche zu vergleichen und zu erkennen, dass das eine Verhalten als angemessener als das andere angesehen wird.
- Der Unterbefehl `TRANSFORMATION` gibt eine lineare Transformation der Ähnlichkeiten mit einem konstanten Term an. Diese Methode ist geeignet, wenn ein Unterschied zwischen den Ähnlichkeiten von 1 Punkt im gesamten Bereich der 10-Punkte-Skala äquivalent ist. Wenn die Lernende ihre Werte so vergeben haben, dass der Unterschied zwischen 0 und 1 dem Unterschied zwischen 5 und 6 entspricht, ist eine lineare Transformation daher geeignet.
- Mit dem Unterbefehl `PLOT` werden Diagramme des gemeinsamen Raums und Transformationsdiagramme angefordert.
- Für alle anderen Parameter gelten die Standardwerte.

Maße

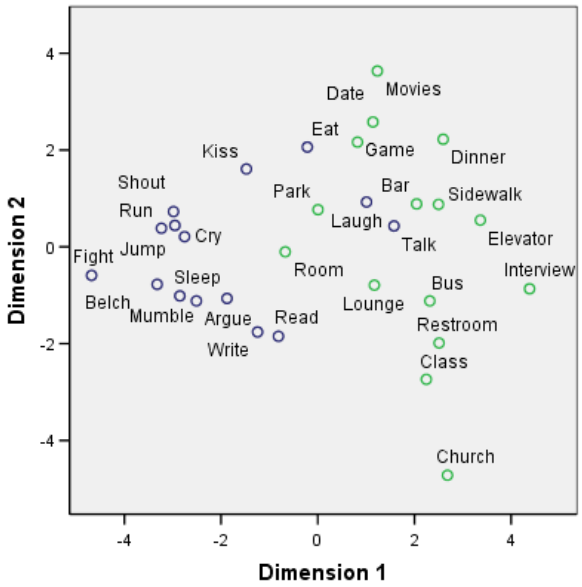
Abbildung 15-28
Maße

Iterationen		169
Endgültiger Funktionswert		.6427725
Wertteile der Funktion	Stressteil	.1900001
	Penalisierungsteil	2.1745069
Fehlende Anpassung	Normalisierter Stress	.0361000
	Stress I nach Kruskal	.1900001
	Stress II nach Kruskal	.5224668
	S-Stress I nach Young	.2760971
	S-Stress II nach Young	.4525933
Güte der Anpassung	Erklärte Streuung	.9639000
	Erklärte Varianz	.8082862
	Ermittelte Bevorzugungsordnungen	.8608333
	Spearman-Rho	.8981120
	Kendall-Tau-b	.7202452
Variationskoeffizienten	Variation der Ähnlichkeiten	.5138436
	Variation der transformierten Ähnlichkeiten	.4751934
	Variation der Distanzen	.3912592
Indizes für Degeneration	Quadratsumme von DeSarbos Indizes für Vermischung	.4957969
	Shepards Index für Nicht-Degeneration	.7173810

Der Algorithmus konvergiert nach 169 Iterationen bei einem endgültigen penalisierten Stress von 0,6427725. Die Variationskoeffizienten und Shepards Index sind ausreichend groß und die Indizes nach DeSarbo ausreichend klein, um den Schluss zuzulassen, dass keine Degenerationsprobleme vorliegen.

Gemeinsamer Raum

Abbildung 15-29
Verbundenes Diagramm des gemeinsamen Raums

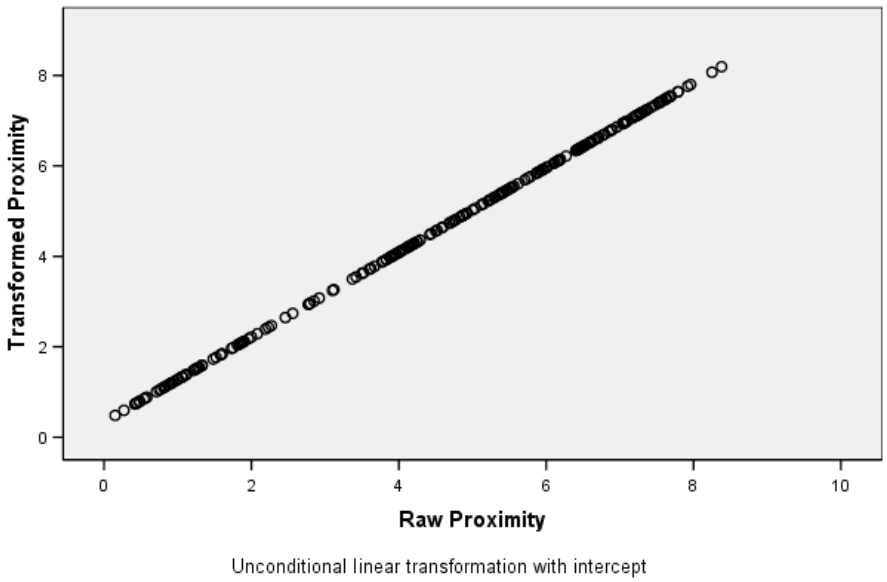


Die horizontale Dimension scheint enger mit den Spaltenobjekten (Verhalten) zusammenzuhängen und unterscheidet zwischen unangemessenem Verhalten (“fighting”, “belching”) und angemessenerem Verhalten. Die vertikale Dimension scheint enger mit den Zeilenobjekten (Situationen) zusammenzuhängen und gibt verschiedene Beschränkungen zwischen Verhalten und Situation vor.

- Am unteren Ende der vertikalen Dimension befinden sich Situationen (“church”, “class”), in denen das Verhalten auf die ruhigeren/introspektiven Verhaltensweisen (“read”, “write”) beschränkt ist. Diese Verhaltensweisen sind daher auf der vertikalen Achse weiter unten zu finden.
- Am oberen Ende der vertikalen Dimension befinden sich Situationen (“movies”, “game”, “date”), in denen das Verhalten auf die sozialen/extrovertierten Verhaltensweisen (“eat”, “kiss”, “laugh”) beschränkt ist. Diese Verhaltensweisen sind daher auf der vertikalen Achse weiter oben zu finden.
- In der Mitte der vertikalen Dimension werden Situationen in der horizontal Dimension entsprechend der allgemeinen Eingeschränktheit der Situation unterteilt. Die von den Verhaltensweise weiter entfernten (“interview”) sind am meisten eingeschränkt, während die näher an den Verhaltensweisen liegenden (“room”, “park”) allgemein weniger eingeschränkt sind.

Ähnlichkeitstransformationen

Abbildung 15-30
Transformationsdiagramm



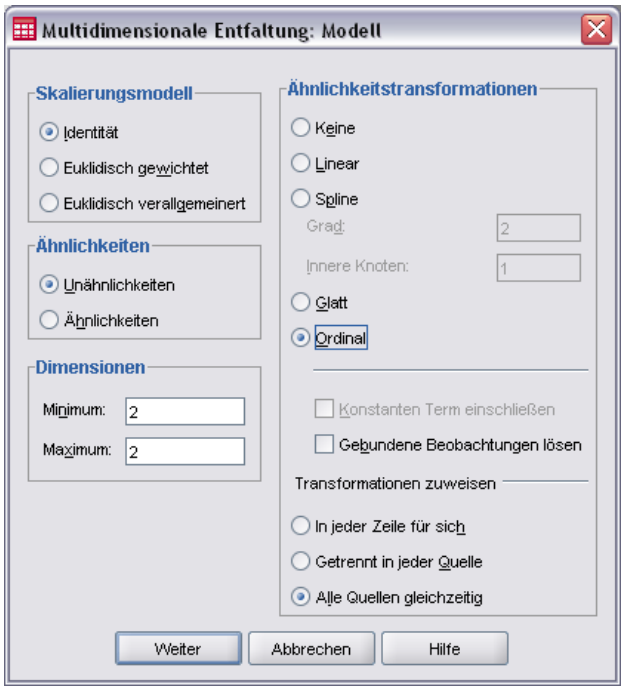
Die Ähnlichkeiten wurden in dieser Analyse als linear behandelt, sodass das Diagramm der transformierten Werte im Vergleich zu den ursprünglichen Ähnlichkeiten eine gerade Linie bildet. Die Anpassungsgüte dieser Lösung ist gut, mit einer anderen Transformation der Ähnlichkeiten kann jedoch möglicherweise eine bessere Anpassung erreicht werden.

Ändern der Ähnlichkeitstransformation (Ordinal)

- Um eine Lösung mit einer ordinalen Transformation der Ähnlichkeiten zu erstellen, klicken Sie auf die Schaltfläche “Zuletzt verwendete Dialogfelder” und wählen Sie “Multidimensionale Entfaltung” aus.

▶ Klicken Sie im Dialogfeld “Multidimensionale Entfaltung” auf Modell.

Abbildung 15-31
Dialogfeld “Modell”



▶ Wählen Sie Ordinal als Ähnlichkeitstransformation.

▶ Klicken Sie auf Weiter.

▶ Klicken Sie im Dialogfeld “Multidimensionale Entfaltung” auf OK.

Im Folgenden finden Sie die durch diese Optionen generierte Befehlssyntax:

```
PREFSCAL
  VARIABLES=Laufen Reden Küssen Schreiben Essen Schlafen Murmeln Lesen Kämpfen
  Rülpsen Diskutieren Springen Weinen Lachen Schreien
  /INPUT=ROWS(ROWID )
  /INITIAL=( 'samplesDirectory/behavior_ini.sav' )
  dim1 dim2
  /CONDITION=UNCONDITIONAL
  /TRANSFORMATION=ORDINAL (KEEPTIES)
  /PROXIMITIES=DISSIMILARITIES
  /MODEL=IDENTITY
  /CRITERIA=DIMENSIONS(2,2) DIFFSTRESS(.000001) MINSTRESS(.0001)
  MAXITER(5000)
  /PENALTY=LAMBDA(0.5) OMEGA(1.0)
  /PRINT=MEASURES COMMON
  /PLOT=COMMON TRANSFORMATIONS .
```

- Die einzige Änderung besteht im Unterbefehl TRANSFORMATION. Die Transformation wurde als ORDINAL festgelegt. Dabei wird die Ordnung der Ähnlichkeiten beibehalten, die transformierten Werte müssen jedoch nicht proportional zu den ursprünglichen Werten sein.

Maße

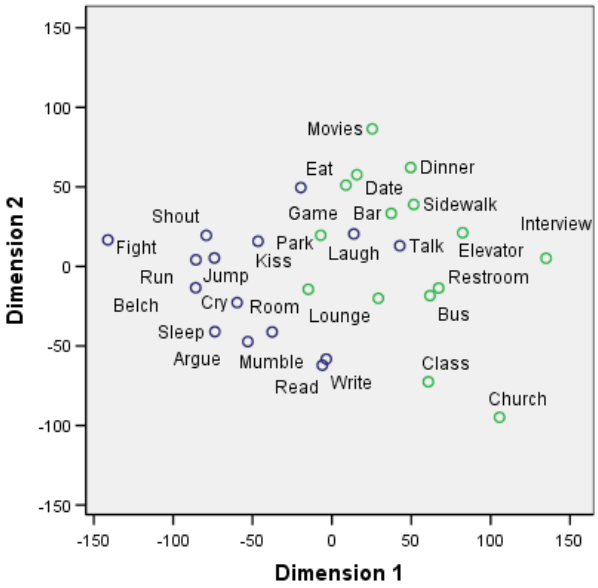
Abbildung 15-32
Maße für Lösung mit ordinaler Transformation

Iterationen		268
Endgültiger Funktionswert		.6044671
Wertteile der Funktion	Stressteil	.1747239
	Penalisierungsteil	2.0911875
Fehlende Anpassung	Normalisierter Stress	.0305285
	Stress I nach Kruskal	.1747239
	Stress II nach Kruskal	.4444641
	S-Stress I nach Young	.2707147
	S-Stress II nach Young	.3978003
Güte der Anpassung	Erklärte Streuung	.9694715
	Erklärte Varianz	.8454488
	Ermittelte Bevorzugungsordnungen	.8574206
	Spearman-Rho	.9032676
	Kendall-Tau-b	.7532788
Variationskoeffizienten	Variation der Ähnlichkeiten	.5138436
	Variation der transformierten Ähnlichkeiten	.4930018
	Variation der Distanzen	.4284849
Indizes für Degeneration	Quadratsumme von DeSarbos Indizes für Vermischung	.3610680
	Shepards Index für Nicht-Degeneration	.7469048

Der Algorithmus konvergiert nach 268 Iterationen bei einem endgültigen penalisierten Stress von 0,6044671. Diese Statistik und die anderen Maße sind für diese Lösung etwas besser geeignet als die entsprechenden Werte bei einer linearen Transformation der Ähnlichkeiten.

Gemeinsamer Raum

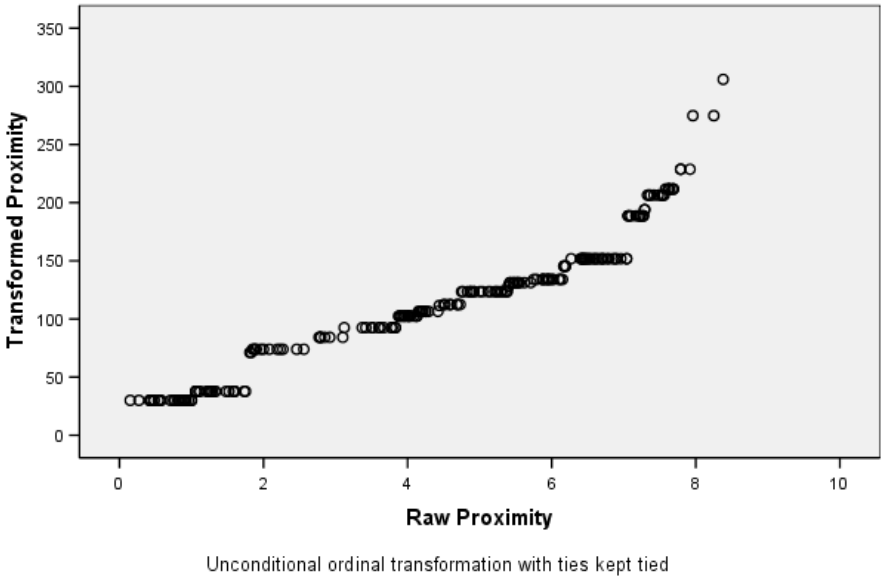
Abbildung 15-33
Verbundenes Diagramm des gemeinsamen Raums für eine Lösung mit ordinaler Transformation



Die Interpretation des gemeinsamen Raums ist bei beiden Lösungen identisch. Möglicherweise weist diese Lösung (mit ordinaler Transformation) relativ geringere Variationen in der vertikalen Dimension als in der horizontalen Dimension auf, als sich in der Lösung mit linearer Transformation zeigt.

Ähnlichkeitstransformationen

Abbildung 15-34
Transformationsdiagramm für eine Lösung mit ordinaler Transformation



Abgesehen von den Werten mit den größten Ähnlichkeiten, die gegenüber den restlichen Variablen aufwärts abweichen, ist die ordinale Transformation der Ähnlichkeiten recht linear. Auf diese Ähnlichkeiten ist wahrscheinlich der größte Teil der Unterschiede zwischen der ordinalen und der linearen Lösung zurückzuführen. Es liegen jedoch nicht genügend Informationen vor, um zu ermitteln, ob es sich bei diesem nicht linearen Trend in den höheren Werten um einen tatsächlichen Trend oder um eine Anomalie handelt.

Empfohlene Literatur

In den folgenden Texten finden Sie weitere Informationen:

Busing, F. M. T. A., P. J. F. Groenen, als auch W. J. Heiser. 2005. Avoiding degeneracy in multidimensional unfolding by penalizing on the coefficient of variation. *Psychometrika*, 70, 71–98.

Green, P. E., als auch V. Rao. 1972. *Applied multidimensional scaling*. Hinsdale, Ill.: Dryden Press.

Price, R. H., als auch D. L. Bouffard. 1974. Behavioral appropriateness and situational constraints as dimensions of social behavior. *Journal of Personality and Social Psychology*, 30, 579–586.

Anhang

A

Beispieldateien

Die zusammen mit dem Produkt installierten Beispieldateien finden Sie im Unterverzeichnis *Samples* des Installationsverzeichnisses.

Beschreibungen

Im Folgenden finden Sie Kurzbeschreibungen der in den verschiedenen Beispielen in der Dokumentation verwendeten Beispieldateien:

- **accidents.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, bei der es um eine Versicherungsgesellschaft geht, die alters- und geschlechtsabhängige Risikofaktoren für Autounfälle in einer bestimmten Region untersucht. Jeder Fall entspricht einer Kreuzklassifikation von Alterskategorie und Geschlecht.
- **adl.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, bei der es um Bemühungen geht, die Vorteile einer vorgeschlagenen Therapieform für Schlaganfallpatienten zu ermitteln. Ärzte teilten weibliche Schlaganfallpatienten nach dem Zufallsprinzip jeweils einer von zwei Gruppen zu. Die erste Gruppe erhielt die physische Standardtherapie, die zweite erhielt eine zusätzliche Emotionaltherapie. Drei Monate nach den Behandlungen wurden die Fähigkeiten der einzelnen Patienten, übliche Alltagsaktivitäten auszuführen, als ordinale Variablen bewertet.
- **advert.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, bei der es um die Bemühungen eines Einzelhändlers geht, die Beziehungen zwischen den in Werbung investierten Beträgen und den daraus resultierenden Umsätzen zu untersuchen. Zu diesem Zweck hat er die Umsätze vergangener Jahre und die zugehörigen Werbeausgaben zusammengestellt.
- **aflatoxin.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, bei der es um Tests von Maisernten auf Aflatoxin geht, ein Gift, dessen Konzentration stark zwischen und innerhalb von Ernteerträgen schwankt. Ein Kornverarbeitungsbetrieb hat aus 8 Ernteerträgen je 16 Proben erhalten und das Aflatoxinniveau in Teilen pro Milliarde (parts per billion, PPB) gemessen.
- **aflatoxin20.sav.** Diese Datendatei enthält die Aflatoxinmessungen aus jeder der 16 Stichproben aus den Erträgen 4 und 8 der Datendatei *aflatoxin.sav*.
- **anorectic.sav.** Bei der Ausarbeitung einer standardisierten Symptomatologie anorektischen/bulimischen Verhaltens führten Forscher (Van der Ham, Meulman, Van Strien, als auch Van Engeland, 1997) eine Studie mit 55 Jugendlichen mit bekannten Ess-Störungen durch. Jeder Patient wurde vier Mal über einen Zeitraum von vier Jahren untersucht, es fanden also insgesamt 220 Beobachtungen statt. Bei jeder Beobachtung erhielten die Patienten Scores für jedes von 16 Symptomen. Die Symptomwerte fehlen für Patient 71

zum Zeitpunkt 2, Patient 76 zum Zeitpunkt 2 und Patient 47 zum Zeitpunkt 3, wodurch 217 gültige Beobachtungen verbleiben.

- **autoaccidents.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, bei der es um die Bemühungen eines Versicherungsanalysten geht, ein Modell zur Anzahl der Autounfälle pro Fahrer unter Berücksichtigung von Alter und Geschlecht zu erstellen. Jeder Fall stellt einen Fahrer dar und erfasst das Geschlecht des Fahrers, sein Alter in Jahren und die Anzahl der Autounfälle in den letzten fünf Jahren.
- **band.sav.** Diese Datendatei enthält die hypothetischen wöchentlichen Verkaufszahlen von CDs für eine Musikgruppe. Daten für drei mögliche Einflussvariablen wurden ebenfalls aufgenommen.
- **bankloan.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, bei der es um die Bemühungen einer Bank geht, den Anteil der nicht zurückgezahlten Kredite zu reduzieren. Die Datei enthält Informationen zum Finanzstatus und demografischen Hintergrund von 850 früheren und potenziellen Kunden. Bei den ersten 700 Fällen handelt es sich um Kunden, denen bereits ein Kredit gewährt wurde. Bei den letzten 150 Fällen handelt es sich um potenzielle Kunden, deren Kreditrisiko die Bank als gering oder hoch einstufen möchte.
- **bankloan_binning.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, die Informationen zum Finanzstatus und demografischen Hintergrund von 5.000 früheren Kunden enthält.
- **behavior.sav.** In einem klassischen Beispiel (Price als auch Bouffard, 1974) wurden 52 Schüler/Studenten gebeten, die Kombinationen aus 15 Situationen und 15 Verhaltensweisen auf einer 10-Punkte-Skala von 0 = "ausgesprochen angemessen" bis 9 = "ausgesprochen unangemessen" zu bewerten. Die Werte werden über die einzelnen Personen gemittelt und als Unähnlichkeiten verwendet.
- **behavior_ini.sav.** Diese Datendatei enthält eine Ausgangskonfiguration für eine zweidimensionale Lösung für *behavior.sav*.
- **brakes.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, bei der es um die Qualitätskontrolle in einer Fabrik geht, die Scheibenbremsen für Hochleistungsautomobile herstellt. Die Datendatei enthält Messungen des Durchmessers von 16 Scheiben aus 8 Produktionsmaschinen. Der Zieldurchmesser für die Scheiben ist 322 Millimeter.
- **breakfast.sav.** In einer klassischen Studie (Green als auch Rao, 1972) wurden 21 MBA-Studenten der Wharton School mit ihren Lebensgefährten darum gebeten, 15 Frühstücksartikel in der Vorzugsreihenfolge von 1 = "am meisten bevorzugt" bis 15 = "am wenigsten bevorzugt" zu ordnen. Die Bevorzugungen wurden in sechs unterschiedlichen Szenarien erfasst, von "Overall preference" (Allgemein bevorzugt) bis "Snack, with beverage only" (Imbiss, nur mit Getränk).
- **breakfast-overall.sav.** Diese Datei enthält die Daten zu den bevorzugten Frühstücksartikeln, allerdings nur für das erste Szenario, "Overall preference" (Allgemein bevorzugt).
- **broadband_1.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, die die Anzahl der Abonnenten eines Breitband-Service, nach Region geordnet, enthält. Die Datendatei enthält die monatlichen Abonnentenzahlen für 85 Regionen über einen Zeitraum von vier Jahren.
- **broadband_2.sav.** Diese Datendatei stimmt mit *broadband_1.sav* überein, enthält jedoch Daten für weitere drei Monate.

- **car_insurance_claims.sav.** Ein an anderer Stelle (McCullagh als auch Nelder, 1989) vorgestelltes und analysiertes Daten-Set bezieht sich auf Schadensansprüche für Autos. Die durchschnittliche Höhe der Schadensansprüche lässt sich mit Gamma-Verteilung modellieren. Dazu wird eine inverse Link-Funktion verwendet, um den Mittelwert der abhängigen Variablen mit einer linearen Kombination aus Alter des Versicherungsnehmers, Fahrzeugtyp und Fahrzeugalter in Bezug zu setzen. Die Anzahl der eingereichten Schadensansprüche kann als Skalierungsgewicht verwendet werden.
- **car_sales.sav.** Diese Datendatei enthält hypothetische Verkaufsschätzer, Listenpreise und physische Spezifikationen für verschiedene Fahrzeugfabrikate und -modelle. Die Listenpreise und physischen Spezifikationen wurden von *edmunds.com* und Hersteller-Websites entnommen.
- **carpet.sav.** In einem beliebten Beispiel möchte (Green als auch Wind, 1973) einen neuen Teppichreiniger vermarkten und dazu den Einfluss von fünf Faktoren auf die Bevorzugung durch den Verbraucher untersuchen: Verpackungsgestaltung, Markenname, Preis, Gütesiegel, *Good Housekeeping* und Geld-zurück-Garantie. Die Verpackungsgestaltung liegt in drei Faktorstufen vor, die sich durch die Position der Auftragebürste unterscheiden. Außerdem gibt es drei Markennamen (*K2R*, *Glory* und *Bissell*), drei Preisstufen sowie je zwei Stufen (Nein oder Ja) für die letzten beiden Faktoren. 10 Kunden stufen 22 Profile ein, die durch diese Faktoren definiert sind. Die Variable *Preference* enthält den Rang der durchschnittlichen Einstufung für die verschiedenen Profile. Ein niedriger Rang bedeutet eine starke Bevorzugung. Diese Variable gibt ein Gesamtmaß der Bevorzugung für die Profile an.
- **carpet_prefs.sav.** Diese Datendatei beruht auf denselben Beispielen, wie für *carpet.sav* beschrieben, enthält jedoch die tatsächlichen Einstufungen durch jeden der 10 Kunden. Die Kunden wurden gebeten, die 22 Produktprofile in der Reihenfolge ihrer Präferenzen einzustufen. Die Variablen *PREF1* bis *PREF22* enthalten die IDs der zugeordneten Profile, wie in *carpet_plan.sav* definiert.
- **catalog.sav.** Diese Datendatei enthält hypothetische monatliche Verkaufszahlen für drei Produkte, die von einem Versandhaus verkauft werden. Daten für fünf mögliche Einflussvariablen wurden ebenfalls aufgenommen.
- **catalog_seasfac.sav.** Diese Datendatei ist mit *catalog.sav* identisch, außer, dass ein Set von saisonalen Faktoren, die mithilfe der Prozedur "Saisonale Zerlegung" berechnet wurden, sowie die zugehörigen Datumsvariablen hinzugefügt wurden.
- **cellular.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, bei der es um die Bemühungen eines Mobiltelefonunternehmens geht, die Kundenabwanderung zu verringern. Scores für die Abwanderungsneigung (von 0 bis 100) werden auf die Kunden angewendet. Kunden mit einem Score von 50 oder höher streben vermutlich einen Anbieterwechsel an.
- **ceramics.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, bei der es um die Bemühungen eines Herstellers geht, der ermitteln möchte, ob ein neue, hochwertige Keramiklegierung eine größere Hitzebeständigkeit aufweist als eine Standardlegierung. Jeder Fall entspricht einem Test einer der Legierungen; die Temperatur, bei der das Keramikwälzlager versagte, wurde erfasst.
- **cereal.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, bei der es um eine Umfrage geht, bei der 880 Personen nach ihren Frühstückgewohnheiten befragt wurden. Außerdem wurden Alter, Geschlecht, Familienstand und Vorliegen bzw. Nichtvorliegen eines aktiven Lebensstils (auf der Grundlage von mindestens zwei Trainingseinheiten pro Woche) erfasst. Jeder Fall entspricht einem Teilnehmer.

- **clothing_defects.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, bei der es um die Qualitätskontrolle in einer Bekleidungsfabrik geht. Aus jeder in der Fabrik produzierten Charge entnehmen die Kontrolleure eine Stichprobe an Bekleidungsartikeln und zählen die Anzahl der Bekleidungsartikel die inakzeptabel sind.
- **coffee.sav.** Diese Datendatei enthält Daten zum wahrgenommenen Image von sechs Eiskaffeemarken (Kennedy, Riquier, als auch Sharp, 1996). Bei den 23 Attributen des Eiskaffee-Image sollten die Teilnehmer jeweils alle Marken auswählen, die durch dieses Attribut beschrieben werden. Die sechs Marken werden als "AA", "BB", "CC", "DD", "EE" und "FF" bezeichnet, um Vertraulichkeit zu gewährleisten.
- **contacts.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, bei der es um die Kontaktlisten einer Gruppe von Vertretern geht, die Computer an Unternehmen verkaufen. Die einzelnen Kontaktpersonen werden anhand der Abteilung, in der sie in ihrem Unternehmen arbeiten und anhand ihrer Stellung in der Unternehmenshierarchie in Kategorien eingeteilt. Außerdem werden der Betrag des letzten Verkaufs, die Zeit seit dem letzten Verkauf und die Größe des Unternehmens, in dem die Kontaktperson arbeitet, aufgezeichnet.
- **creditpromo.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, bei der es um die Bemühungen eines Kaufhauses geht, die Wirksamkeit einer kürzlich durchgeführten Kreditkarten-Werbeaktion einzuschätzen. Dazu wurden 500 Karteninhaber nach dem Zufallsprinzip ausgewählt. Die Hälfte erhielt eine Werbebeilage, die einen reduzierten Zinssatz für Einkäufe in den nächsten drei Monaten ankündigte. Die andere Hälfte erhielt eine Standard-Werbebeilage.
- **customer_dbase.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, bei der es um die Bemühungen eines Unternehmens geht, das die Informationen in seinem Data Warehouse nutzen möchte, um spezielle Angebote für Kunden zu erstellen, die mit der größten Wahrscheinlichkeit darauf ansprechen. Nach dem Zufallsprinzip wurde eine Untergruppe des Kundenstamms ausgewählt. Diese Gruppe erhielt die speziellen Angebote und die Reaktionen wurden aufgezeichnet.
- **customers_model.sav.** Diese Datei enthält hypothetische Daten zu Einzelpersonen, auf die sich eine Marketingkampagne richtete. Zu diesen Daten gehören demografische Informationen, eine Übersicht über die bisherigen Einkäufe und die Angabe ob die einzelnen Personen auf die Kampagne ansprachen oder nicht. Jeder Fall entspricht einer Einzelperson.
- **customers_new.sav.** Diese Datei enthält hypothetische Daten zu Einzelpersonen, die potenzielle Kandidaten für Marketingkampagnen sind. Zu diesen Daten gehören demografische Informationen und eine Übersicht über die bisherigen Einkäufe für jede Person. Jeder Fall entspricht einer Einzelperson.
- **debate.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, die gepaarte Antworten auf eine Umfrage unter den Zuhörern einer politischen Debatte enthält (Antworten vor und nach der Debatte). Jeder Fall entspricht einem Befragten.
- **debate_aggregate.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, in der die Antworten aus *debate.sav* aggregiert wurden. Jeder Fall entspricht einer Kreuzklassifikation der bevorzugten Politiker vor und nach der Debatte.
- **demo.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, bei der es um eine Kundendatenbank geht, die zum Zwecke der Zusendung monatlicher Angebote erworben wurde. Neben verschiedenen demografischen Informationen ist erfasst, ob der Kunde auf das Angebot geantwortet hat.

- **demo_cs_1.sav.** Hierbei handelt es sich um eine hypothetische Datendatei für den ersten Schritt eines Unternehmens, das eine Datenbank mit Umfrageinformationen zusammenstellen möchte. Jeder Fall entspricht einer anderen Stadt. Außerdem sind IDs für Region, Provinz, Landkreis und Stadt erfasst.
- **demo_cs_2.sav.** Hierbei handelt es sich um eine hypothetische Datendatei für den zweiten Schritt eines Unternehmens, das eine Datenbank mit Umfrageinformationen zusammenstellen möchte. Jeder Fall entspricht einem anderen Stadtteil aus den im ersten Schritt ausgewählten Städten. Außerdem sind IDs für Region, Provinz, Landkreis, Stadt, Stadtteil und Wohneinheit erfasst. Die Informationen zur Stichprobenziehung aus den ersten beiden Stufen des Stichprobenplans sind ebenfalls enthalten.
- **demo_cs.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, die Umfrageinformationen enthält die mit einem komplexen Stichprobenplan erfasst wurden. Jeder Fall entspricht einer anderen Wohneinheit. Es sind verschiedene Informationen zum demografischen Hintergrund und zur Stichprobenziehung erfasst.
- **dietstudy.sav.** Diese hypothetische Datendatei enthält die Ergebnisse einer Studie der "Stillman-Diät" (Rickman, Mitchell, Dingman, als auch Dalen, 1974). Jeder Fall entspricht einem Teilnehmer und enthält dessen Gewicht vor und nach der Diät in amerikanischen Pfund sowie mehrere Messungen des Triglyceridspiegels (in mg/100 ml).
- **dischargedata.sav.** Hierbei handelt es sich um eine Datendatei zum Thema *Seasonal Patterns of Winnipeg Hospital Use*, (Menec , Roos, Nowicki, MacWilliam, Finlayson , als auch Black, 1999) (Saisonale Muster der Belegung im Krankenhaus von Winnipeg) vom Manitoba Centre for Health Policy.
- **dvdplayer.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, bei der es um die Entwicklung eines neuen DVD-Spielers geht. Mithilfe eines Prototyps hat das Marketing-Team Zielgruppendaten erfasst. Jeder Fall entspricht einem befragten Benutzer und enthält demografische Daten zu dem Benutzer sowie dessen Antworten auf Fragen zum Prototyp.
- **flying.sav.** Diese Datendatei enthält die Flugmeilen zwischen zehn Städten in den USA.
- **german_credit.sav.** Diese Daten sind aus dem Daten-Set "German credit" im Repository of Machine Learning Databases (Blake als auch Merz, 1998) an der Universität von Kalifornien in Irvine entnommen.
- **grocery_1month.sav.** Bei dieser hypothetischen Datendatei handelt es sich um die Datendatei *grocery_coupons.sav*, wobei die wöchentlichen Einkäufe zusammengefasst sind, sodass jeder Fall einem anderen Kunden entspricht. Dadurch entfallen einige der Variablen, die wöchentlichen Änderungen unterworfen waren, und der verzeichnete ausgegebene Betrag ist nun die Summe der Beträge, die in den vier Wochen der Studie ausgegeben wurden.
- **grocery_coupons.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, die Umfragedaten enthält, die von einer Lebensmittelkette erfasst wurden, die sich für die Kaufgewohnheiten ihrer Kunden interessiert. Jeder Kunde wird über vier Wochen beobachtet, und jeder Fall entspricht einer Kundenwoche und enthält Informationen zu den Geschäften, in denen der Kunde einkauft sowie zu anderen Merkmalen, beispielsweise welcher Betrag in der betreffenden Woche für Lebensmittel ausgegeben wurde.
- **guttman.sav.** Bell (Bell, 1961) legte eine Tabelle zur Darstellung möglicher sozialer Gruppen vor. Guttman (Guttman, 1968) verwendete einen Teil dieser Tabelle, bei der fünf Variablen, die Aspekte beschreiben, wie soziale Interaktion, das Gefühl der Gruppenzugehörigkeit, die

physische Nähe der Mitglieder und die Formalität der Beziehung, mit sieben theoretischen sozialen Gruppen gekreuzt wurden: “crowds” (Menschenmassen, beispielsweise die Zuschauer eines Fußballspiels), “audience” (Zuhörerschaften, beispielsweise die Personen im Theater oder bei einer Vorlesung), “public” (Öffentlichkeit, beispielsweise Zeitungsleser oder Fernsehzuschauer), “mobs” (Mobs, wie Menschenmassen, jedoch mit wesentlich stärkerer Interaktion), “primary groups” (Primärgruppen, vertraulich), “secondary groups” (Sekundärgruppen, freiwillig) und “modern community” (die moderne Gesellschaft, ein lockerer Zusammenschluss, der aus einer engen physischen Nähe und dem Bedarf an spezialisierten Dienstleistungen entsteht).

- **healthplans.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, bei der es um die Bemühungen einer Versicherungsgruppe geht, vier verschiedene Pläne zur Gesundheitsvorsorge für Kleinbetriebe zu evaluieren. Zwölf Inhaber von Kleinbetrieben (Arbeitgeber) wurden gebeten, die Pläne danach in eine Rangfolge zu bringen, wie gern sie sie ihren Mitarbeitern anbieten würden. Jeder Fall entspricht einem Arbeitgeber und enthält die Reaktionen auf die einzelnen Pläne.
- **health_funding.sav.** Hierbei handelt es sich um eine hypothetische Datei, die Daten zur Finanzierung des Gesundheitswesens (Betrag pro 100 Personen), Krankheitsraten (Rate pro 10.000 Personen der Bevölkerung) und Besuche bei medizinischen Einrichtungen/Ärzten (Rate pro 10.000 Personen der Bevölkerung) enthält. Jeder Fall entspricht einer anderen Stadt.
- **hivassay.sav.** Hierbei handelt es sich um eine hypothetische Datendatei zu den Bemühungen eines pharmazeutischen Labors, einen Schnelltest zur Erkennung von HIV-Infektionen zu entwickeln. Die Ergebnisse des Tests sind acht kräftiger werdende Rotschattierungen, wobei kräftigeren Schattierungen auf eine höhere Infektionswahrscheinlichkeit hindeuten. Bei 2.000 Blutproben, von denen die Hälfte mit HIV infiziert war, wurde ein Labortest durchgeführt.
- **hourlywagedata.sav.** Hierbei handelt es sich um eine hypothetische Datendatei zum Stundenlohn von Pflegepersonal in Praxen und Krankenhäusern mit unterschiedlich langer Berufserfahrung.
- **insure.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, bei der es um eine Versicherungsgesellschaft geht, die die Risikofaktoren untersucht, die darauf hinweisen, ob ein Kunde die Leistungen einer mit einer Laufzeit von 10 Jahren abgeschlossenen Lebensversicherung in Anspruch nehmen wird. Jeder Fall in der Datendatei entspricht einem Paar von Verträgen, je einer mit Leistungsforderung und der andere ohne, wobei die beiden Versicherungsnehmer in Alter und Geschlecht übereinstimmen.
- **judges.sav.** Hierbei handelt es sich um eine hypothetische Datendatei mit den Wertungen von ausgebildeten Kampfrichtern (sowie eines Sportliebhabers) zu 300 Kunstturnleistungen. Jede Zeile stellt eine Leistung dar; die Kampfrichter bewerteten jeweils dieselben Leistungen.
- **kinship_dat.sav.** Rosenberg und Kim (Rosenberg als auch Kim, 1975) haben 15 Bezeichnungen für den Verwandtschaftsgrad untersucht (Tante, Bruder, Cousin, Tochter, Vater, Enkelin, Großvater, Großmutter, Enkel, Mutter, Neffe, Nichte, Schwester, Sohn, Onkel). Die beiden Analytiker baten vier Gruppen von College-Studenten (zwei weibliche und zwei männliche Gruppen), diese Bezeichnungen auf der Grundlage der Ähnlichkeiten zu sortieren. Zwei Gruppen (eine weibliche und eine männliche Gruppe) wurden gebeten, die Bezeichnungen zweimal zu sortieren; die zweite Sortierung sollte dabei nach einem anderen Kriterium erfolgen als die erste. So wurden insgesamt sechs “Quellen” erzielt. Jede Quelle entspricht einer Ähnlichkeitsmatrix mit 15×15 Elementen. Die Anzahl der Zellen ist dabei gleich der

Anzahl der Personen in einer Quelle minus der Anzahl der gemeinsamen Platzierungen der Objekte in dieser Quelle.

- **kinship_ini.sav.** Diese Datendatei enthält eine Ausgangskonfiguration für eine dreidimensionale Lösung für *kinship_dat.sav*.
- **kinship_var.sav.** Diese Datendatei enthält die unabhängigen Variablen *gender* (Geschlecht), *gener* (Generation) und *degree* (Verwandtschaftsgrad), die zur Interpretation der Dimensionen einer Lösung für *kinship_dat.sav* verwendet werden können. Insbesondere können sie verwendet werden, um den Lösungsraum auf eine lineare Kombination dieser Variablen zu beschränken.
- **mailresponse.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, in der es um die Bemühungen eines Bekleidungsherstellers geht, der ermitteln möchte, ob die Verwendung von Briefsendungen für das Direktmarketing zu schnelleren Antworten führt als Postwurfsendungen. Die Mitarbeiter in der Bestellannahme erfassen, wie vielen Wochen nach der Postsendung die einzelnen Bestellungen aufgegeben wurden.
- **marketvalues.sav.** Diese Datendatei betrifft Hausverkäufe in einem Neubaugebiet in Algonquin, Illinois, in den Jahren 1999–2000. Diese Verkäufe sind in Grundbucheinträgen dokumentiert.
- **mutualfund.sav.** Diese Datendatei betrifft Aktienmarktdaten für verschiedene Technologieaktien, die in im Index S&P 500 verzeichnet sind. Jeder Fall entspricht einem Unternehmen.
- **nhis2000_subset.sav.** Die "National Health Interview Survey (NHIS)" ist eine große, bevölkerungsbezogene Umfrage in unter der US-amerikanischen Zivilbevölkerung. Es werden persönliche Interviews in einer landesweit repräsentativen Stichprobe von Haushalten durchgeführt. Für die Mitglieder jedes Haushalts werden demografische Informationen und Beobachtungen zum Gesundheitsverhalten und Gesundheitsstatus eingeholt. Diese Datendatei enthält eine Teilmenge der Informationen aus der Umfrage des Jahres 2000. National Center for Health Statistics. National Health Interview Survey, 2000. Datendatei und Dokumentation öffentlich zugänglich. *ftp://ftp.cdc.gov/pub/Health_Statistics/NCHS/Datasets/NHIS/2000/*. Zugriff erfolgte 2003.
- **ozone.sav.** Die Daten enthalten 330 Beobachtungen zu sechs meteorologischen Variablen zur Vorhersage der Ozonkonzentration aus den übrigen Variablen. Bei früheren Untersuchungen (Breiman als auch Friedman, 1985), (Hastie als auch Tibshirani, 1990) fanden Wissenschaftler einige Nichtlinearitäten unter diesen Variablen, die die Standardverfahren bei der Regression behindern.
- **pain_medication.sav.** Diese hypothetische Datendatei enthält die Ergebnisse eines klinischen Tests für ein entzündungshemmendes Medikament zur Schmerzbehandlung bei chronischer Arthritis. Von besonderem Interesse ist die Zeitdauer, bis die Wirkung des Medikaments einsetzt und wie es im Vergleich mit bestehenden Medikamenten abschneidet.
- **patient_los.sav.** Diese hypothetische Datendatei enthält die Behandlungsaufzeichnungen zu Patienten, die wegen des Verdachts auf Herzinfarkt in das Krankenhaus eingeliefert wurden. Jeder Fall entspricht einem Patienten und enthält diverse Variablen in Bezug auf den Krankenhausaufenthalt.

- **patlos_sample.sav.** Diese hypothetische Datendatei enthält die Behandlungsaufzeichnungen für eine Stichprobe von Patienten, denen während der Behandlung eines Herzinfarkts Thrombolytika verabreicht wurden. Jeder Fall entspricht einem Patienten und enthält diverse Variablen in Bezug auf den Krankenhausaufenthalt.
- **polishing.sav.** Hierbei handelt es sich um die Datendatei "Nambeware Polishing Times" aus der Data and Story Library. Sie bezieht sich auf die Bemühungen eines Herstellers von Metallgeschirr (Nambe Mills, Santa Fe, New Mexico) zur zeitlichen Planung seiner Produktion. Jeder Fall entspricht einem anderen Artikel in der Produktpalette. Für jeden Artikel sind Durchmesser, Polierzeit, Preis und Produkttyp erfasst.
- **poll_cs.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, bei der es um Bemühungen geht, die öffentliche Unterstützung für einen Gesetzentwurf zu ermitteln, bevor er im Parlament eingebracht wird. Die Fälle entsprechen registrierten Wählern. Für jeden Fall sind County, Gemeinde und Wohnviertel des Wählers erfasst.
- **poll_cs_sample.sav.** Diese hypothetische Datendatei enthält eine Stichprobe der in *poll_cs.sav* aufgeführten Wähler. Die Stichprobe wurde gemäß dem in der Plandatei *poll.csplan* angegebenen Stichprobenplan gezogen und in dieser Datendatei sind die Einschlusswahrscheinlichkeiten und Stichprobengewichtungen erfasst. Beachten Sie jedoch Folgendes: Da im Stichprobenplan die PPS-Methode (PPS: probability proportional to size; Wahrscheinlichkeit proportional zur Größe) verwendet wird, gibt es außerdem eine Datei mit den gemeinsamen Auswahlwahrscheinlichkeiten (*poll_jointprob.sav*). Die zusätzlichen Variablen zum demografischen Hintergrund der Wähler und ihrer Meinung zum vorgeschlagenen Gesetzentwurf wurden nach der Ziehung der Stichprobe erfasst und zur Datendatei hinzugefügt.
- **property_assess.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, in der es um die Bemühungen eines für einen Bezirk (County) zuständigen Immobilienbewerters geht, trotz eingeschränkter Ressourcen die Einschätzungen des Werts von Immobilien auf dem aktuellsten Stand zu halten. Die Fälle entsprechen den Immobilien, die im vergangenen Jahr in dem betreffenden County verkauft wurden. Jeder Fall in der Datendatei enthält die Gemeinde, in der sich die Immobilie befindet, den Bewerter, der die Immobilie besichtigt hat, die seit dieser Bewertung verstrichene Zeit, den zu diesem Zeitpunkt ermittelten Wert sowie den Verkaufswert der Immobilie.
- **property_assess_cs.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, in der es um die Bemühungen eines für einen US-Bundesstaat zuständigen Immobilienbewerters geht, trotz eingeschränkter Ressourcen die Einschätzungen des Werts von Immobilien auf dem aktuellsten Stand zu halten. Die Fälle entsprechen den Immobilien in dem betreffenden Bundesstaat. Jeder Fall in der Datendatei enthält das County, die Gemeinde und das Wohnviertel, in dem sich die Immobilie befindet, die seit der letzten Bewertung verstrichene Zeit sowie zu diesem Zeitpunkt ermittelten Wert.
- **property_assess_cs_sample.sav.** Diese hypothetische Datendatei enthält eine Stichprobe der in *property_assess_cs.sav* aufgeführten Immobilien. Die Stichprobe wurde gemäß dem in der Plandatei *property_assess.csplan* angegebenen Stichprobenplan gezogen und in dieser Datendatei sind die Einschlusswahrscheinlichkeiten und Stichprobengewichtungen erfasst. Die zusätzliche Variable *Current value* (Aktueller Wert) wurde nach der Ziehung der Stichprobe erfasst und zur Datendatei hinzugefügt.

- **recidivism.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, bei der es um die Bemühungen einer Strafverfolgungsbehörde geht, einen Einblick in die Rückfallraten in ihrem Zuständigkeitsbereich zu gewinnen. Jeder Fall entspricht einem frühren Straftäter und erfasst Daten zu dessen demografischen Hintergrund, einige Details zu seinem ersten Verbrechen sowie die Zeit bis zu seiner zweiten Festnahme, sofern diese innerhalb von zwei Jahren nach der ersten Festnahme erfolgte.
- **recidivism_cs_sample.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, bei der es um die Bemühungen einer Strafverfolgungsbehörde geht, einen Einblick in die Rückfallraten in ihrem Zuständigkeitsbereich zu gewinnen. Jeder Fall entspricht einem früheren Straftäter, der im Juni 2003 erstmals aus der Haft entlassen wurde, und erfasst Daten zu dessen demografischen Hintergrund, einige Details zu seinem ersten Verbrechen sowie die Daten zu seiner zweiten Festnahme, sofern diese bis Ende Juni 2006 erfolgte. Die Straftäter wurden aus per Stichprobenziehung ermittelten Polizeidirektionen ausgewählt (gemäß dem in *recidivism_cs.csplan* angegebenen Stichprobenplan). Da hierbei eine PPS-Methode (PPS: probability proportional to size; Wahrscheinlichkeit proportional zur Größe) verwendet wird, gibt es außerdem eine Datei mit den gemeinsamen Auswahlwahrscheinlichkeiten (*recidivism_cs_jointprob.sav*).
- **salesperformance.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, bei der es um Bewertung von zwei neuen Verkaufsschulungen geht. 60 Mitarbeiter, die in drei Gruppen unterteilt sind, erhalten jeweils eine Standardschulung. Zusätzlich erhält Gruppe 2 eine technische Schulung und Gruppe 3 eine Praxisschulung. Die einzelnen Mitarbeiter wurden am Ende der Schulung einem Test unterzogen und die erzielten Punkte wurden erfasst. Jeder Fall in der Datendatei stellt einen Lehrgangsteilnehmer dar und enthält die Gruppe, der der Lehrgangsteilnehmer zugeteilt wurde sowie die von ihm in der Prüfung erreichte Punktzahl.
- **satisf.sav.** Hierbei handelt es sich um eine hypothetische Datendatei zu einer Zufriedenheitsumfrage, die von einem Einzelhandelsunternehmen in 4 Filialen durchgeführt wurde. Insgesamt wurden 582 Kunden befragt. Jeder Fall gibt die Antworten eines einzelnen Kunden wieder.
- **screws.sav.** Diese Datendatei enthält Informationen zu den Eigenschaften von Schrauben, Bolzen, Muttern und Reißnägeln (Hartigan, 1975).
- **shampoo_ph.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, bei der es um die Qualitätskontrolle in einer Fabrik für Haarpflegeprodukte geht. In regelmäßigen Zeitabständen werden Messwerte von sechs separaten Ausgangschargen erhoben und ihr pH-Wert erfasst. Der Zielbereich ist 4,5–5,5.
- **ships.sav.** Ein an anderer Stelle (McCullagh et al., 1989) vorgestelltes und analysiertes Daten-Set bezieht sich auf die durch Wellen verursachten Schäden an Frachtschiffen. Die Vorfallshäufigkeiten können unter Angabe von Schiffstyp, Konstruktionszeitraum und Betriebszeitraum gemäß einer Poisson-Rate modelliert werden. Das Aggregat der Betriebsmonate für jede Zelle der durch die Kreuzklassifizierung der Faktoren gebildeten Tabelle gibt die Werte für die Risikoanfälligkeit an.
- **site.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, bei der es um die Bemühungen eines Unternehmens geht, neue Standorte für die betriebliche Expansion auszuwählen. Das Unternehmen beauftragte zwei Berater unabhängig voneinander mit der Bewertung der Standorte. Neben einem umfassenden Bericht gaben die Berater auch eine zusammenfassende Wertung für jeden Standort als “good” (gut) “fair” (mittelmäßig) oder “poor” (schlecht) ab.

- **siteratings.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, bei der es um die Betatests der neuen Website eines E-Commerce-Unternehmens geht. Jeder Fall entspricht einem Beta-Tester, der die Brauchbarkeit der Website auf einer Skala von 0 bis 20 bewertete.
- **smokers.sav.** Diese Datendatei wurde aus der Umfrage "National Household Survey of Drug Abuse" aus dem Jahr 1998 abstrahiert und stellt eine Wahrscheinlichkeitsstichprobe US-amerikanischer Haushalte dar. Daher sollte der erste Schritt bei der Analyse dieser Datendatei darin bestehen, die Daten entsprechend den Bevölkerungstrends zu gewichten.
- **smoking.sav.** Hierbei handelt es sich um eine von Greenacre (Greenacre , 1984) vorgestellte hypothetische Tabelle. Die relevante Tabelle wird durch eine Kreuztabelle der Rauchgewohnheiten und der Berufskategorie gebildet. Die Variable *Berufsgruppe* enthält die Berufskategorien *Senior Manager*, *Junior Manager*, *Angestellter mit Erfahrung*, *Angestellter ohne Erfahrung* und *Sekretariat* sowie die Kategorie *National Average*, die als Ergänzung der Analyse dienen kann. Die Variable *Rauchen* enthält die Rauchgewohnheiten *Nichtraucher*, *Leicht*, *Mittel* und *Stark* sowie die Kategorien *No Alcohol* und *Alcohol*, die als Ergänzung der Analyse dienen können.
- **storebrand.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, bei der es um die Bemühungen einer Verkaufsleiterin in einem Lebensmittelmarkt geht, die die Verkaufszahlen des Waschmittels der Eigenmarke gegenüber den anderen Marken steigern möchte. Sie erarbeitet eine Werbeaktion im Geschäft und spricht an der Kasse mit Kunden. Jeder Fall entspricht einem Kunden.
- **stores.sav.** Diese Datendatei enthält hypothetische monatliche Marktanteilsdaten für zwei konkurrierende Lebensmittelgeschäfte. Jeder Fall entspricht den Marktanteilsdaten für einen bestimmten Monat.
- **stroke_clean.sav.** Diese hypothetische Datendatei enthält den Zustand einer medizinischen Datenbank, nachdem diese mithilfe der Prozeduren in der Option "Data Preparation" bereinigt wurde.
- **stroke_invalid.sav.** Diese hypothetische Datendatei enthält den ursprünglichen Zustand einer medizinischen Datenbank, der mehrere Dateneingabefehler aufweist.
- **stroke_survival.** In dieser hypothetischen Datendatei geht es um die Überlebenszeiten von Patienten, die nach einem Rehabilitationsprogramm wegen eines ischämischen Schlaganfalls mit einer Reihe von Problemen zu kämpfen haben. Nach dem Schlaganfall werden das Auftreten von Herzinfarkt, ischämischem Schlaganfall und hämorrhagischem Schlaganfall sowie der Zeitpunkt des Ereignisses aufgezeichnet. Die Stichprobe ist auf der linken Seite abgeschnitten, da sie nur Patienten enthält, die bis zum Ende des Rehabilitationprogramms, das nach dem Schlaganfall durchgeführt wurde, überlebten.
- **stroke_valid.sav.** Diese hypothetische Datendatei enthält den Zustand einer medizinischen Datenbank, nachdem diese mithilfe der Prozedur "Daten validieren" überprüft wurde. Sie enthält immer noch potenziell anomale Fälle.
- **tastetest.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, bei der es um Bewertung der Auswirkungen der Mulchfarbe auf den Geschmack von Pflanzenprodukten geht. Der Geschmack von Erdbeeren, die in rotem, blauem und schwarzem Rindenmulch gezogen wurden, wurde von Testpersonen auf einer ordinalen Skala (weit unter bis weit über dem Durchschnitt) bewertet. Jeder Fall entspricht einem Geschmackstester.

- **telco.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, bei der es um die Bemühungen eines Telekommunikationsunternehmens geht, die Kundenabwanderung zu verringern. Jeder Fall entspricht einem Kunden und enthält verschiedene Informationen zum demografischen Hintergrund und zur Servicenutzung.
- **telco_extra.sav.** Diese Datendatei ähnelt der Datei *telco.sav*, allerdings wurden die Variablen "tenure" und die Log-transformierten Variablen zu den Kundenausgaben entfernt und durch standardisierte Log-transformierte Variablen ersetzt.
- **telco_missing.sav.** Diese Datendatei entspricht der Datei *telco_mva_complete.sav*, allerdings wurde ein Teil der Daten durch fehlende Werte ersetzt.
- **telco_mva_complete.sav.** Bei dieser Datendatei handelt es sich um eine Teilmenge der Datendatei *telco.sav*, allerdings mit anderen Variablennamen.
- **testmarket.sav.** Diese hypothetische Datendatei bezieht sich auf die Pläne einer Fast-Food-Kette, einen neuen Artikel in ihr Menü aufzunehmen. Es gibt drei mögliche Kampagnen zur Verkaufsförderung für das neue Produkt. Daher wird der neue Artikel in Filialen in mehreren zufällig ausgewählten Märkten eingeführt. An jedem Standort wird eine andere Form der Verkaufsförderung verwendet und die wöchentlichen Verkaufszahlen für das neue Produkt werden für die ersten vier Wochen aufgezeichnet. Jeder Fall entspricht einer Standort-Woche.
- **testmarket_1month.sav.** Bei dieser hypothetischen Datendatei handelt es sich um die Datendatei *testmarket.sav*, wobei die wöchentlichen Verkaufszahlen zusammengefasst sind, sodass jeder Fall einem Standort entspricht. Dadurch entfallen einige der Variablen, die wöchentlichen Änderungen unterworfen waren, und die verzeichneten Verkaufszahlen sind nun die Summe der Verkaufszahlen während der vier Wochen der Studie.
- **tree_car.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, die demografische Daten sowie Daten zum Kaufpreis von Fahrzeugen enthält.
- **tree_credit.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, die demografische Daten sowie Daten zu früheren Bankkrediten enthält.
- **tree_missing_data.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, die demografische Daten sowie Daten zu früheren Bankkrediten enthält und eine große Anzahl fehlender Werte aufweist.
- **tree_score_car.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, die demografische Daten sowie Daten zum Kaufpreis von Fahrzeugen enthält.
- **tree_textdata.sav.** Eine einfache Datendatei mit nur zwei Variablen, die vor allem den Standardzustand von Variablen vor der Zuweisung von Messniveau und Wertelabels zeigen soll.
- **tv-survey.sav.** Hierbei handelt es sich um eine hypothetische Datendatei zu einer Studie, die von einem Fernsehstudio durchgeführt wurde, das überlegt, ob die Laufzeit eines erfolgreichen Programms verlängert werden soll. 906 Personen wurden gefragt, ob sie das Programm unter verschiedenen Bedingungen ansehen würden. Jede Zeile entspricht einem Befragten; jede Spalte entspricht einer Bedingung.
- **ulcer_recurrence.sav.** Diese Datei enthält Teilinformationen aus einer Studie zum Vergleich der Wirksamkeit zweier Therapien zur Vermeidung des Wiederauftretens von Geschwüren. Es stellt ein gutes Beispiel für intervallzensierte Daten dar und wurde an anderer Stelle (Collett, 2003) vorgestellt und analysiert.

- **ulcer_recurrence_recoded.sav.** In dieser Datei sind die Daten aus *ulcer_recurrence.sav* so umstrukturiert, dass das Modell der Ereigniswahrscheinlichkeit für jedes Intervall der Studie berechnet werden kann und nicht nur die Ereigniswahrscheinlichkeit am Ende der Studie. Sie wurde an anderer Stelle (Collett et al., 2003) vorgestellt und analysiert.
- **verd1985.sav.** Diese Datendatei enthält eine Umfrage (Verdegaal, 1985). Die Antworten von 15 Subjekten auf 8 Variablen wurden aufgezeichnet. Die relevanten Variablen sind in drei Sets unterteilt. Set 1 umfasst *alter* und *heirat*, Set 2 besteht aus *pet* und *news* und in Set 3 finden sich *music* und *live*. Die Variable *pet* wird mehrfach nominal skaliert und die Variable *Alter* ordinal. Alle anderen Variablen werden einzeln nominal skaliert.
- **virus.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, bei der es um die Bemühungen eines Internet-Dienstanbieters geht, der die Auswirkungen eines Virus auf seine Netzwerke ermitteln möchte. Dabei wurde vom Moment der Virusentdeckung bis zu dem Zeitpunkt, zu dem die Virusinfektion unter Kontrolle war, der (ungefähre) prozentuale Anteil infizierter E-Mail in den Netzwerken erfasst.
- **waittimes.sav.** Hierbei handelt es sich um eine hypothetische Datendatei zu den Wartezeiten für Kunden bei drei verschiedenen Filialen einer Bank. Jeder Fall entspricht einem Kunden und zeichnet die Wartezeit und die Filiale.
- **webusability.sav.** Hierbei handelt es sich um eine hypothetische Datendatei, bei der es um Tests zur Benutzerfreundlichkeit eines neuen Internetgeschäfts geht. Jeder Fall entspricht einer von fünf Testpersonen, die die Benutzerfreundlichkeit bewerten und gibt für sechs separate Aufgaben an, ob die Testperson sie erfolgreich ausführen könnte.
- **wheeze_steubenville.sav.** Hierbei handelt es sich um eine Teilmenge der Daten aus einer Langzeitstudie zu den gesundheitlichen Auswirkungen der Luftverschmutzung auf Kinder (Ware, Dockery, Spiro III, Speizer, als auch Ferris Jr., 1984). Die Daten enthalten wiederholte binäre Messungen des Keuchens von Kindern aus Steubenville, Ohio, im Alter von 7, 8, 9 und 10 Jahren sowie eine unveränderlichen Angabe, ob die Mutter im ersten Jahr der Studie rauchte oder nicht.
- **workprog.sav.** Hierbei handelt es sich um eine hypothetische Datendatei zu einem Arbeitsprogramm der Regierung, das versucht, benachteiligten Personen bessere Arbeitsplätze zu verschaffen. Eine Stichprobe potenzieller Programmteilnehmer wurde beobachtet. Von diesen Personen wurden nach dem Zufallsprinzip einige für die Teilnahme an dem Programm ausgewählt. Jeder Fall entspricht einem Programmteilnehmer.

Bibliografie

Barlow, R. E., D. J. Bartholomew, D. J. Bremner, als auch H. D. Brunk. 1972. *Statistical inference under order restrictions*. New York: John Wiley and Sons.

Bell, E. H. 1961. *Social foundations of human behavior: Introduction to the study of sociology*. New York: Harper & Row.

Benzécri, J. P. 1969. Statistical analysis as a tool to make patterns emerge from data. In: *Methodologies of Pattern Recognition,* S. Watanabe (Hg.). New York: Academic Press, 35–74.

Benzécri, J. P. 1992. *Correspondence analysis handbook*. New York: Marcel Dekker.

Bishop, Y. M., S. E. Feinberg, als auch P. W. Holland. 1975. *Discrete multivariate analysis: Theory and practice*. Cambridge, Massachusetts: MIT Press.

Blake, C. L., als auch C. J. Merz. 1998. "UCI Repository of machine learning databases." Available at http://www.ics.uci.edu/~mlearn/MLRepository.html.

Breiman, L., als auch J. H. Friedman. 1985. Estimating optimal transformations for multiple regression and correlation. *Journal of the American Statistical Association*, 80, 580–598.

Buja, A. 1990. Remarks on functional canonical variates, alternating least squares methods and ACE. *Annals of Statistics*, 18, 1032–1069.

Busing, F. M. T. A., P. J. F. Groenen, als auch W. J. Heiser. 2005. Avoiding degeneracy in multidimensional unfolding by penalizing on the coefficient of variation. *Psychometrika*, 70, 71–98.

Carroll, J. D. 1968. Generalization of canonical correlation analysis to three or more sets of variables. In: *Proceedings of the 76th Annual Convention of the American Psychological Association, 3,* Washington, D.C.: American Psychological Association, 227–228.

Collett, D. 2003. *Modelling survival data in medical research*, 2 (Hg.). Boca Raton: Chapman & Hall/CRC.

Commandeur, J. J. F., als auch W. J. Heiser. 1993. *Mathematical derivations in the proximity scaling (PROXSCAL) of symmetric data matrices*. Leiden: Department of Data Theory, Universität Leiden.

De Haas, M., J. A. Algera, H. F. J. M. Van Tuijl, als auch J. J. Meulman. 2000. Macro and micro goal setting: In search of coherence. *Applied Psychology*, 49, 579–595.

De Leeuw, J. 1982. Nonlinear principal components analysis. In: *COMPSTAT Proceedings in Computational Statistics,* Wien: Physica Verlag, 77–89.

De Leeuw, J. 1984. *Canonical analysis of categorical data*, 2nd (Hg.). Leiden: DSWO Press.

De Leeuw, J. 1984. The Gifi system of nonlinear multivariate analysis. In: *Data Analysis and Informatics III,* E. Diday, et al. (Hg.)., 415–424.

De Leeuw, J. 1990. Multivariate analysis with optimal scaling. In: *Progress in Multivariate Analysis,* S. Das Gupta, als auch J. Sethuraman (Hgg.). Kalkutta: Indian Statistical Institute.

De Leeuw, J., als auch W. J. Heiser. 1980. Multidimensional scaling with restrictions on the configuration. In: *Multivariate Analysis, Vol. V,* P. R. Krishnaiah (Hg.). Amsterdam: North-Holland, 501–522.

De Leeuw, J., als auch J. Van Rijckevorsel. 1980. HOMALS and PRINCALS—Some generalizations of principal components analysis. In: *Data Analysis and Informatics,* E. Diday, et al. (Hg.). Amsterdam: North-Holland, 231–242.

De Leeuw, J., F. W. Young, als auch Y. Takane. 1976. Additive structure in qualitative data: An alternating least squares method with optimal scaling features. *Psychometrika*, 41, 471–503.

Eckart, C., als auch G. Young. 1936. The approximation of one matrix by another one of lower rank. *Psychometrika*, 1, 211–218.

Fisher, R. A. 1938. *Statistical methods for research workers*. Edinburgh: Oliver and Boyd.

Fisher, R. A. 1940. The precision of discriminant functions. *Annals of Eugenics*, 10, 422–429.

Gabriel, K. R. 1971. The biplot graphic display of matrices with application to principal components analysis. *Biometrika*, 58, 453–467.

Gifi, A. 1985. *PRINCALS. Research Report UG-85-02*. Leiden: Department of Data Theory, Universität Leiden.

Gifi, A. 1990. *Nonlinear multivariate analysis*. Chichester: John Wiley and Sons.

Gilula, Z., als auch S. J. Haberman. 1988. The analysis of multivariate contingency tables by restricted canonical and restricted association models. *Journal of the American Statistical Association*, 83, 760–771.

Gower, J. C., als auch J. J. Meulman. 1993. The treatment of categorical information in physical anthropology. *International Journal of Anthropology*, 8, 43–51.

Green, P. E., als auch V. Rao. 1972. *Applied multidimensional scaling*. Hinsdale, Ill.: Dryden Press.

Green, P. E., als auch Y. Wind. 1973. *Multiattribute decisions in marketing: A measurement approach*. Hinsdale, Ill.: Dryden Press.

Greenacre , M. J. 1984. *Theory and applications of correspondence analysis*. London: Academic Press.

Guttman, L. 1941. The quantification of a class of attributes: A theory and method of scale construction. In: *The Prediction of Personal Adjustment,* P. Horst (Hg.). New York: Social Science Research Council, 319–348.

Guttman, L. 1968. A general nonmetric technique for finding the smallest coordinate space for configurations of points. *Psychometrika*, 33, 469–506.

Hartigan, J. A. 1975. *Clustering algorithms*. New York: John Wiley and Sons.

Hastie, T., als auch R. Tibshirani. 1990. *Generalized additive models*. London: Chapman and Hall.

Hastie, T., R. Tibshirani, als auch A. Buja. 1994. Flexible discriminant analysis. *Journal of the American Statistical Association*, 89, 1255–1270.

Hayashi, C. 1952. On the prediction of phenomena from qualitative data and the quantification of qualitative data from the mathematico-statistical point of view. *Annals of the Institute of Statitical Mathematics*, 2, 93–96.

Heiser, W. J. 1981. *Unfolding analysis of proximity data*. Leiden: Department of Data Theory, Universität Leiden.

Heiser, W. J., als auch F. M. T. A. Busing. 2004. Multidimensional scaling and unfolding of symmetric and asymmetric proximity relations. In: *Handbook of Quantitative Methodology for the Social Sciences,* D. Kaplan (Hg.). Thousand Oaks, Kalifornien: Sage Publications, Inc, 25–48.

Heiser, W. J., als auch J. J. Meulman. 1994. Homogeneity analysis: Exploring the distribution of variables and their nonlinear relationships. In: *Correspondence Analysis in the Social Sciences: Recent Developments and Applications,* M. Greenacre, als auch J. Blasius (Hgg.). New York: Academic Press, 179–209.

Heiser, W. J., als auch J. J. Meulman. 1995. Nonlinear methods for the analysis of homogeneity and heterogeneity. In: *Recent Advances in Descriptive Multivariate Analysis,* W. J. Krzanowski (Hg.). Oxford: Oxford University Press, 51–89.

Horst, P. 1961. Generalized canonical correlations and their applications to experimental data. *Journal of Clinical Psychology*, 17, 331–347.

Horst, P. 1961. Relations among m sets of measures. *Psychometrika*, 26, 129–149.

Israëls, A. 1987. *Eigenvalue techniques for qualitative data*. Leiden: DSWO Press.

Kennedy, R., C. Riquier, als auch B. Sharp. 1996. Practical applications of correspondence analysis to categorical data in market research. *Journal of Targeting, Measurement, and Analysis for Marketing*, 5, 56–70.

Kettenring, J. R. 1971. Canonical analysis of several sets of variables. *Biometrika*, 58, 433–460.

Kruskal, J. B. 1964. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika*, 29, 1–28.

Kruskal, J. B. 1964. Nonmetric multidimensional scaling: A numerical method. *Psychometrika*, 29, 115–129.

Kruskal, J. B. 1965. Analysis of factorial experiments by estimating monotone transformations of the data. *Journal of the Royal Statistical Society Series B*, 27, 251–263.

Kruskal, J. B. 1978. Factor analysis and principal components analysis: Bilinear methods. In: *International Encyclopedia of Statistics,* W. H. Kruskal, als auch J. M. Tanur (Hgg.). New York: The Free Press, 307–330.

Kruskal, J. B., als auch R. N. Shepard. 1974. A nonmetric variety of linear factor analysis. *Psychometrika*, 39, 123–157.

Krzanowski, W. J., als auch F. H. C. Marriott. 1994. *Multivariate analysis: Part I, distributions, ordination and inference*. London: Edward Arnold.

Lebart, L., A. Morineau, als auch K. M. Warwick. 1984. *Multivariate descriptive statistical analysis*. New York: John Wiley and Sons.

Lingoes, J. C. 1968. The multivariate analysis of qualitative data. *Multivariate Behavioral Research*, 3, 61–94.

Max, J. 1960. Quantizing for minimum distortion. *Proceedings IEEE (Information Theory)*, 6, 7–12.

McCullagh, P., als auch J. A. Nelder. 1989. *Generalized Linear Models*, 2nd (Hg.). London: Chapman & Hall.

Menec , V., N. Roos, D. Nowicki, L. MacWilliam, G. Finlayson , als auch C. Black. 1999. *Seasonal Patterns of Winnipeg Hospital Use*. : Manitoba Centre for Health Policy.

Meulman, J. J., A. J. Van der Kooij, als auch A. Babinec. 2000. New features of categorical principal components analysis for complicated data sets, including data mining. In: *Classification, Automation and New Media,* W. Gaul, als auch G. Ritter (Hgg.). Berlin: Springer-Verlag, 207–217.

Meulman, J. J., A. J. Van der Kooij, als auch W. J. Heiser. 2004. Principal components analysis with nonlinear optimal scaling transformations for ordinal and nominal data. In: *Handbook of Quantitative Methodology for the Social Sciences,* D. Kaplan (Hg.). Thousand Oaks, Kalifornien: Sage Publications, Inc, 49–70.

Meulman, J. J. 1982. *Homogeneity analysis of incomplete data*. Leiden: DSWO Press.

Meulman, J. J. 1986. *A distance approach to nonlinear multivariate analysis*. Leiden: DSWO Press.

Meulman, J. J. 1992. The integration of multidimensional scaling and multivariate analysis with optimal transformations of the variables. *Psychometrika*, 57, 539–565.

Meulman, J. J. 1993. Principal coordinates analysis with optimal transformations of the variables: Minimizing the sum of squares of the smallest eigenvalues. *British Journal of Mathematical and Statistical Psychology*, 46, 287–300.

Meulman, J. J. 1996. Fitting a distance model to homogeneous subsets of variables: Points of view analysis of categorical data. *Journal of Classification*, 13, 249–266.

Meulman, J. J. 2003. Prediction and classification in nonlinear data analysis: Something old, something new, something borrowed, something blue. *Psychometrika*, 4, 493–517.

Meulman, J. J., als auch W. J. Heiser. 1997. Graphical display of interaction in multiway contingency tables by use of homogeneity analysis. In: *Visual Display of Categorical Data,* M. Greenacre, als auch J. Blasius (Hgg.). New York: Academic Press, 277–296.

Meulman, J. J., als auch P. Verboon. 1993. Points of view analysis revisited: Fitting multidimensional structures to optimal distance components with cluster restrictions on the variables. *Psychometrika*, 58, 7–35.

Nishisato, S. 1980. *Analysis of categorical data: Dual scaling and its applications*. Toronto: University of Toronto Press.

Nishisato, S. 1984. Forced classification: A simple application of a quantification method. *Psychometrika*, 49, 25–36.

Nishisato, S. 1994. *Elements of dual scaling: An introduction to practical data analysis*. Hillsdale, N.J.: Lawrence Erlbaum Associates, Inc.

Pratt, J. W. 1987. Dividing the indivisible: Using simple symmetry to partition variance explained. In: *Proceedings of the Second International Conference in Statistics,* T. Pukkila, als auch S. Puntanen (Hgg.). Tampere, Finnland: Universität Tampere, 245–260.

Price, R. H., als auch D. L. Bouffard. 1974. Behavioral appropriateness and situational constraints as dimensions of social behavior. *Journal of Personality and Social Psychology*, 30, 579–586.

Ramsay, J. O. 1989. Monotone regression splines in action. *Statistical Science*, 4, 425–441.

Rao, C. R. 1973. *Linear statistical inference and its applications*, 2 (Hg.). New York: John Wiley and Sons.

Rao, C. R. 1980. Matrix approximations and reduction of dimensionality in multivariate statistical analysis. In: *Multivariate Analysis, Vol. 5,* P. R. Krishnaiah (Hg.). Amsterdam: North-Holland, 3–22.

Rickman, R., N. Mitchell, J. Dingman, als auch J. E. Dalen. 1974. Changes in serum cholesterol during the Stillman Diet. *Journal of the American Medical Association*, 228, 54–58.

Rosenberg, S., als auch M. P. Kim. 1975. The method of sorting as a data-gathering procedure in multivariate research. *Multivariate Behavioral Research*, 10, 489–502.

Roskam, E. E. 1968. *Metric analysis of ordinal data in psychology*. Voorschoten: VAM.

Shepard, R. N. 1962. The analysis of proximities: Multidimensional scaling with an unknown distance function I. *Psychometrika*, 27, 125–140.

Shepard, R. N. 1962. The analysis of proximities: Multidimensional scaling with an unknown distance function II. *Psychometrika*, 27, 219–246.

Shepard, R. N. 1966. Metric structures in ordinal data. *Journal of Mathematical Psychology*, 3, 287–315.

Tenenhaus, M., als auch F. W. Young. 1985. An analysis and synthesis of multiple correspondence analysis, optimal scaling, dual scaling, homogeneity analysis, and other methods for quantifying categorical multivariate data. *Psychometrika*, 50, 91–119.

Theunissen, N. C. M., J. J. Meulman, A. L. Den Ouden, H. M. Koopman, G. H. Verrips, S. P. Verloove-Vanhorick, als auch J. M. Wit. 2003. Changes can be studied when the measurement instrument is different at different time points. *Health Services and Outcomes Research Methodology*, 4, 109–126.

Tucker, L. R. 1960. Intra-individual and inter-individual multidimensionality. In: *Psychological Scaling: Theory & Applications,* H. Gulliksen, als auch S. Messick (Hgg.). New York: John Wiley and Sons, 155–167.

Van der Burg, E. 1988. *Nonlinear canonical correlation and some related techniques*. Leiden: DSWO Press.

Van der Burg, E., als auch J. De Leeuw. 1983. Nonlinear canonical correlation. *British Journal of Mathematical and Statistical Psychology*, 36, 54–80.

Van der Burg, E., J. De Leeuw, als auch R. Verdegaal. 1988. Homogeneity analysis with k sets of variables: An alternating least squares method with optimal scaling features. *Psychometrika*, 53, 177–197.

Van der Ham, T., J. J. Meulman, D. C. Van Strien, als auch H. Van Engeland. 1997. Empirically based subgrouping of eating disorders in adolescents: A longitudinal perspective. *British Journal of Psychiatry*, 170, 363–368.

Van der Kooij, A. J., als auch J. J. Meulman. 1997. MURALS: Multiple regression and optimal scaling using alternating least squares. In: *Softstat '97,* F. Faulbaum, als auch W. Bandilla (Hgg.). Stuttgart: Gustav Fischer, 99–106.

Van Rijckevorsel, J. 1987. *The application of fuzzy coding and horseshoes in multiple correspondence analysis*. Leiden: DSWO Press.

Verboon, P., als auch I. A. Van der Lans. 1994. Robust canonical discriminant analysis. *Psychometrika*, 59, 485–507.

Verdegaal, R. 1985. *Meer sets analyse voor kwalitatieve gegevens (in niederländischer Sprache)*. Leiden: Department of Data Theory, Universität Leiden.

Vlek, C., als auch P. J. Stallen. 1981. Judging risks and benefits in the small and in the large. *Organizational Behavior and Human Performance*, 28, 235–271.

Wagenaar, W. A. 1988. *Paradoxes of gambling behaviour*. London: Lawrence Erlbaum Associates, Inc.

Ware, J. H., D. W. Dockery, A. Spiro III, F. E. Speizer, als auch B. G. Ferris Jr.. 1984. Passive smoking, gas cooking, and respiratory health of children living in six cities. *American Review of Respiratory Diseases*, 129, 366–374.

Winsberg, S., als auch J. O. Ramsay. 1980. Monotonic transformations to additivity using splines. *Biometrika*, 67, 669–674.

Winsberg, S., als auch J. O. Ramsay. 1983. Monotone spline transformations for dimension reduction. *Psychometrika*, 48, 575–595.

Wolter, K. M. 1985. *Introduction to variance estimation*. Berlin: Springer-Verlag.

Young, F. W. 1981. Quantitative analysis of qualitative data. *Psychometrika*, 46, 357–387.

Young, F. W., J. De Leeuw, als auch Y. Takane. 1976. Regression with qualitative and quantitative variables: An alternating least squares method with optimal scaling features. *Psychometrika*, 41, 505–528.

Young, F. W., Y. Takane, als auch J. De Leeuw. 1978. The principal components of mixed measurement level multivariate data: An alternating least squares method with optimal scaling features. *Psychometrika*, 43, 279–281.

Zeijl, E., Y. te Poel, M. du Bois-Reymond, J. Ravesloot, als auch J. J. Meulman. 2000. The role of parents and peers in the leisure activities of young adolescents. *Journal of Leisure Research*, 32, 281–302.

Index

Ähnlichkeitstransformationen
in “Multidimensionale Entfaltung”, 87
Allgemeines euklidisches Modell
in “Multidimensionale Entfaltung”, 87
Anfängliche Diagramme des gemeinsamen Raums
in “Multidimensionale Entfaltung”, 91
ANOVA
in Kategoriale Regression, 21
Anpassung
in Nichtlineare kanonische Korrelationsanalyse, 44
Anpassungswerte
in Nichtlineare kanonische Korrelationsanalyse, 200
Ausgangskonfiguration
in Kategoriale Regression, 21
in “Multidimensionale Entfaltung”, 89
in Multidimensionale Skalierung, 79
in Nichtlineare kanonische Korrelationsanalyse, 44
Ausreißer
bei der Mehrfachkorrespondenzanalyse, 266

Beispieldateien
Speicherort, 326
Beiträge
in der Korrespondenzanalyse, 225, 239
Biplots
bei der Mehrfachkorrespondenzanalyse, 65
in der Kategorialen Hauptkomponentenanalyse, 36
in der Korrespondenzanalyse, 53, 222

Cronbachs Alpha
in der Kategorialen Hauptkomponentenanalyse, 147

DeSarbos Indizes für Vermischung
in “Multidimensionale Entfaltung”, 293, 296, 302, 309, 319
Deskriptive Statistiken
in Kategoriale Regression, 21
Diagramme
in der Korrespondenzanalyse, 53
in Kategoriale Regression, 23
in Multidimensionale Skalierung, 81–82
in Nichtlineare kanonische Korrelationsanalyse, 44
Diagramme des gemeinsamen Raums
in “Multidimensionale Entfaltung”, 91
in Multidimensionale Skalierung, 81
Diagramme des individuellen Raums
in “Multidimensionale Entfaltung”, 91
in Multidimensionale Skalierung, 81
Diagramme für mehrere Starts
in “Multidimensionale Entfaltung”, 91
Diagramme für Raumgewichtungen
in “Multidimensionale Entfaltung”, 91
Diagramme mit individuellen Raumgewichtungen
in “Multidimensionale Entfaltung”, 91
in Multidimensionale Skalierung, 81
Diagramme mit projizierten Zentroiden
in der Kategorialen Hauptkomponentenanalyse, 37
Dimensionen
in der Korrespondenzanalyse, 50, 238
Dimensionsgewichtungen
in “Multidimensionale Entfaltung”, 304, 311
Diskretisierung
bei der Mehrfachkorrespondenzanalyse, 59
in der Kategorialen Hauptkomponentenanalyse, 29
in Kategoriale Regression, 19
Diskriminationsmaß-Diagramme
bei der Mehrfachkorrespondenzanalyse, 66
Diskriminationsmaße
bei der Mehrfachkorrespondenzanalyse, 63, 260
Distanzen
in “Multidimensionale Entfaltung”, 93
in Multidimensionale Skalierung, 83
Distanzmaße
in der Korrespondenzanalyse, 50
Drei-fach-Entfaltung
in “Multidimensionale Entfaltung”, 297

Eigenwerte
in der Kategorialen Hauptkomponentenanalyse, 147, 153, 170
in Nichtlineare kanonische Korrelationsanalyse, 200
Einschränkungen
in Multidimensionale Skalierung, 78
Einschränkungen im gemeinsamen Raum.
in “Multidimensionale Entfaltung”, 88
Endgültige Diagramme des gemeinsamen Raums
in “Multidimensionale Entfaltung”, 91

Fehlende Werte
bei der Mehrfachkorrespondenzanalyse, 60
in der Kategorialen Hauptkomponentenanalyse, 30
in Kategoriale Regression, 20

Gelockerte Aktualisierungen
in Multidimensionale Skalierung, 79
Gemeinsame Raumkoordinaten
in “Multidimensionale Entfaltung”, 93
in Multidimensionale Skalierung, 83
Gemeinsamer Raum
in “Multidimensionale Entfaltung”, 294, 297, 303, 310, 320, 324

in Multidimensionale Skalierung, 284, 287
Gewichtetes euklidisches Modell
in “Multidimensionale Entfaltung”, 87
Gewichtungen
in Nichtlineare kanonische Korrelationsanalyse, 44, 201

Identitätsmodell
in “Multidimensionale Entfaltung”, 87
Individuelle Raumgewichtungen
in “Multidimensionale Entfaltung”, 93
in Multidimensionale Skalierung, 83
Individuelle Raumkoordinaten
in “Multidimensionale Entfaltung”, 93
Individueller Raum
in “Multidimensionale Entfaltung”, 304, 311
Interkorrelationen
in Kategoriale Regression, 109
Iterationskriterien
in “Multidimensionale Entfaltung”, 89
in Multidimensionale Skalierung, 79
Iterationsprotokoll
bei der Mehrfachkorrespondenzanalyse, 63
in der Kategorialen Hauptkomponentenanalyse, 33, 147
in “Multidimensionale Entfaltung”, 93
in Multidimensionale Skalierung, 83

Kategoriale Hauptkomponentenanalyse, 25, 31, 141, 154
Iterationsprotokoll, 147
Kategoriepunkte, 174
Komponentenladungen, 151, 154, 171
Messniveau der optimalen Skalierung, 27
Modellzusammenfassung, 147, 153, 170
Objektwerte, 150, 153, 172
Quantifizierungen, 148, 168
Variablen speichern, 35
zusätzliche Funktionen beim Befehl, 39
Kategoriale Regression, 16, 97
Anpassungsgüte des Modells, 110
Diagramme, 16
Interkorrelationen, 109
Korrelationen, 110–111
Messniveau der optimalen Skalierung, 17
Residuen, 115
speichern, 23
Statistiken, 16
Transformationsdiagramme, 113
Wichtigkeit, 111
zusätzliche Funktionen beim Befehl, 24
Kategoriekoordinaten
in Nichtlineare kanonische Korrelationsanalyse, 206
Kategoriendiagramme
bei der Mehrfachkorrespondenzanalyse, 66
in der Kategorialen Hauptkomponentenanalyse, 37
Kategorienquantifikationen
bei der Mehrfachkorrespondenzanalyse, 63, 261
in der Kategorialen Hauptkomponentenanalyse, 33
in Kategoriale Regression, 21
in Nichtlineare kanonische Korrelationsanalyse, 44
Kategoriepunkte
in der Kategorialen Hauptkomponentenanalyse, 174
Koeffizienten
in Kategoriale Regression, 110
Komponentenladungen
in der Kategorialen Hauptkomponentenanalyse, 33, 151, 154, 171
in Nichtlineare kanonische Korrelationsanalyse, 44, 203
Komponentenladungsdiagramme
in der Kategorialen Hauptkomponentenanalyse, 38
Konfidenzstatistik
in der Korrespondenzanalyse, 52, 227
Korrelationen
in Multidimensionale Skalierung, 83
Korrelationen nullter Ordnung
in Kategoriale Regression, 111
Korrelationsdiagramme
in Multidimensionale Skalierung, 81
Korrelationsmatrix
bei der Mehrfachkorrespondenzanalyse, 63
in der Kategorialen Hauptkomponentenanalyse, 33
Korrespondenzanalyse, 47, 49–50, 52–53, 216, 218, 234
Beiträge, 225, 239
Biplots, 222
Diagramme, 47
Dimensionen, 238
Konfidenzstatistik, 227
Korrespondenztabellen, 221, 251
Normalisierung, 217
Permutationen, 227
Profile, 223
Spaltenwertdiagramme, 241
Statistiken, 47
Trägheit pro Dimension, 221
Zeilen- und Spaltenwerte, 225
Zeilenwertdiagramme, 241, 252
zusätzliche Funktionen beim Befehl, 55
Korrespondenztabellen
in der Korrespondenzanalyse, 221, 251

Mehrfachkorrespondenzanalyse, 56, 61, 254
Ausreißer, 266
Diskriminationsmaße, 260
Kategorienquantifikationen, 261
Messniveau der optimalen Skalierung, 58
Modellzusammenfassung, 258
Objektwerte, 259, 263
Variablen speichern, 64
zusätzliche Funktionen beim Befehl, 67
Messniveau der optimalen Skalierung
bei der Mehrfachkorrespondenzanalyse, 58
in der Kategorialen Hauptkomponentenanalyse, 27
Modellzusammenfassung
bei der Mehrfachkorrespondenzanalyse, 258
Multidimensionale Entfaltung, 85, 290, 313
Ähnlichkeitstransformationen, 321, 325

Ausgabe, 93
degenerierte Lösungen, 290
Diagramme, 85, 91
Drei-fach-Entfaltung , 297
Einschränkungen im gemeinsamen Raum., 88
Gemeinsamer Raum, 294, 297, 303, 310, 320, 324
Individueller Raum, 304, 311
Maße, 293, 296, 302, 309, 319, 323
Modell, 87
Optionen, 89
Statistiken, 85
zusätzliche Funktionen beim Befehl, 95
Multidimensionale Skalierung, 69, 72–76, 271
Ausgabe, 83
Diagramme, 69, 81–82
Einschränkungen, 78
Gemeinsamer Raum, 284, 287
Modell, 77
Optionen, 79
Statistiken, 69
Stressmaße, 283, 287
Transformationsdiagramme, 286
zusätzliche Funktionen beim Befehl, 84
Multiples *R*
in Kategoriale Regression, 21

Nichtlineare kanonische Korrelationsanalyse, 40, 43–44, 192
Diagramme, 40
Gewichtungen, 201
Kategoriekoordinaten, 206
Komponentenladungen, 201, 203
Quantifizierungen, 204
Statistiken, 40
Zentroide, 207
Zusammenfassung der Analyse, 200
zusätzliche Funktionen beim Befehl, 45
Normalisierung
in der Korrespondenzanalyse, 50, 217

Objektpunktdiagramme
bei der Mehrfachkorrespondenzanalyse, 65
in der Kategorialen Hauptkomponentenanalyse, 36
Objektwerte
bei der Mehrfachkorrespondenzanalyse, 63, 259, 263
in der Kategorialen Hauptkomponentenanalyse, 33, 150, 153, 172
in Nichtlineare kanonische Korrelationsanalyse, 44

Partielle Korrelationen
in Kategoriale Regression, 111
Penalisierter Stress
in “Multidimensionale Entfaltung”, 293, 302, 309, 319, 323
Penalisierungsterm
in “Multidimensionale Entfaltung”, 89
Permutationen
in der Korrespondenzanalyse, 227
PREFSCAL, 85
Prinzipal-Normalisierung
in der Korrespondenzanalyse, 217
Profile
in der Korrespondenzanalyse, 223
Projizierte Zentroide
in Nichtlineare kanonische Korrelationsanalyse, 207

Quantifizierungen
in der Kategorialen Hauptkomponentenanalyse, 148, 168
in Nichtlineare kanonische Korrelationsanalyse, 204

R^2
in Kategoriale Regression, 110
Regressionskoeffizienten
in Kategoriale Regression, 21
Residuen
in Kategoriale Regression, 115
Residuen-Diagramme
in “Multidimensionale Entfaltung”, 91

Shepard-Diagramme
in “Multidimensionale Entfaltung”, 91
Shepards Index für Nicht-Degeneration
in “Multidimensionale Entfaltung”, 293, 296, 302, 309, 319
Skalierungsmodell
in “Multidimensionale Entfaltung”, 87
Spaltenprinzipal-Normalisierung
in der Korrespondenzanalyse, 217
Spaltenwertdiagramme
in der Korrespondenzanalyse, 241
Spaltenwerte
in der Korrespondenzanalyse, 225
Standardisierung
in der Korrespondenzanalyse, 50
Stress-Diagramme
in “Multidimensionale Entfaltung”, 91
in Multidimensionale Skalierung, 81
Stressmaße
in “Multidimensionale Entfaltung”, 93
in Multidimensionale Skalierung, 83, 283, 287
Streudiagramm der Anpassung
in “Multidimensionale Entfaltung”, 91
Symmetrische Normalisierung
in der Korrespondenzanalyse, 217

Teilkorrelationen
in Kategoriale Regression, 111
Trägheit
in der Korrespondenzanalyse, 52, 221, 225
Transformationsdiagramme
bei der Mehrfachkorrespondenzanalyse, 66

in der Kategorialen Hauptkomponentenanalyse, 37
in Kategoriale Regression, 113
in “Multidimensionale Entfaltung”, 91, 321, 325
in Multidimensionale Skalierung, 81, 286
Transformierte Ähnlichkeiten
in “Multidimensionale Entfaltung”, 93
in Multidimensionale Skalierung, 83
Transformierte unabhängige Variablen
in Multidimensionale Skalierung, 83
Triplots
in der Kategorialen Hauptkomponentenanalyse, 36

Variablengewichtung
bei der Mehrfachkorrespondenzanalyse, 58
in der Kategorialen Hauptkomponentenanalyse, 27
Varianz berücksichtigt für
in der Kategorialen Hauptkomponentenanalyse, 33, 147, 170
Variationskoeffizient
in “Multidimensionale Entfaltung”, 293, 296, 302, 309, 319
Verbundene Kategoriendiagramme
bei der Mehrfachkorrespondenzanalyse, 66
in der Kategorialen Hauptkomponentenanalyse, 37
Verbundenes Diagramm der individuellen Räume
in “Multidimensionale Entfaltung”, 304, 311
Verbundenes Diagramm des gemeinsamen Raums
in “Multidimensionale Entfaltung”, 294, 297, 303, 310, 320, 324
Verlustwerte
in Nichtlineare kanonische Korrelationsanalyse, 200

Wichtigkeit
in Kategoriale Regression, 111

Zeilenprinzipal-Normalisierung
in der Korrespondenzanalyse, 217
Zeilenwertdiagramme
in der Korrespondenzanalyse, 241, 252
Zeilenwerte
in der Korrespondenzanalyse, 225
Zentroide
in Nichtlineare kanonische Korrelationsanalyse, 44, 207
Zusatzobjekte
in Kategoriale Regression, 21
Zusatzpunkte
in der Korrespondenzanalyse, 228